AD

Target Acquisition: Human Observer Performance Studies and TARGAC Model Validation

FINAL TECHNICAL REPORT

by

Dr. J.M. Valeton and Dr. P. Bijl

June 1994

United States Army

EUROPEAN RESEARCH OFFICE OF THE US ARMY

London England

CONTRACT NUMBER DAJA45-91-C-0013

TNO Human Factcrs Research Institute, Soesterberg, The Netherlands

Approved for public release; distribution unlimited

REPORT DOCUMENTATION PAGE

ERO Proposal Number: **R&D 6720-EN-01** Contract Number: **DAJA45-91-C-0013**

Tittle of Proposal: **Evaluation of Target Acquisition Models, Part A**

Report Number: **Final** Period Covered: **From Sept. 1991 to Nov. 1993**

Name of Institution:
TNO Human Factors Research Institute
PO Box 23
3769 ZG Soesterberg
The Netherlands

Principal Investigator: **Dr. J.M. Valeton**

ABSTRACT

Human target acquisition performance was studied using thermal imagery that was collected during the field trial BEST TWO, organized by NATO AC243/Panel4/RSG.15 in 1990. In a number of carefully controlled laboratory experiments, recognition and identification probabilities were measured for a large number of stationary and moving targets at ranges between 1 and 4 km. Target detection was not investigated.

The target acquisition model TARGAC was validated by comparing its predictions with observed recognition ranges. For all trials that were used in the observer experiments, TARGAC predictions were calculated on the basis of meteorological, target, background and time information that was measured in the field.

The major conclusion of the observer experiments is that the human acquisition performance depends considerably on factors such as target structure, local terrain structure, and cognitive factors. In TARGAC, these factors are not modelled. For the BEST TWO situation, the model predictions were determined solely by target size and thermal imager resolution limit.

A quantitative comparison between actual and predicted recognition ranges shows that TARGAC systematically underestimates human acquisition performance: on average, observed recognition ranges are a factor of 1.8 longer than the model predictions. Further, the model does not make accurate predictions for individual targets on specific backgrounds: the ratio between observed and predicted recognition range varies between 0.9 to 3.6 (95% criterion), i.e. a factor of 4. The routine that TARGAC uses to calculate electro-optical and human visual system performance is based on the widely-used NVESD Static Performance Model. Hence, the present findings may also be relevant for the validity of other models.

A number of software errors found in TARGAC are documented; most problems have been fixed.

KEYWORDS

Target acquisition
Target acquisition models
Thermal infrared
Observer performance
Target recognition
Target identification

CONTENTS Page

CHAPTER 1

GENERAL OVERVIEW AND CONCLUSIONS

1. MODELS AND VALIDATION

Target acquisition models predict how well human observers, using an optical or electro-optical (E/O) viewing device, are able to detect, recognize or identify military targets. The input variables are the properties of the targets and backgrounds, the atmospheric conditions, and the physical properties of the viewing device used. The output is the probability of correct detection, recognition or identification as a function of target range. Target acquisition models are used, for example, in tactical decision aids (TDA's), in war games, and as a tool to compare performance of competing sensor systems for a specific task. A comprehensive target acquisition model is TARGAC, developed at the U.S. Army Research Laboratory, Battlefield Environment Directorate (ARL-BED). The model is part of the Electro Optical Systems Atmospheric Effects Library (EOSAEL).

Target recognition and identification by human observers is a complex pattern recognition process, that is not yet fully understood. Models that describe this human capability must therefore still rely on a number of simplifying assumptions. In addition to the human pattern recognition capabilities, psychological factors play an important role in human performance, and these factors are even more difficult to incorporate into a model. The quality of these models is therefore not always known. To allow correct and valid use of target acquisition models it is necessary to know how well they predict target acquisition performance, and under which conditions a model may be used. Furthermore, it is important to know the confidence limits of the predicted model output. Also, the fact that the accuracy of the model predictions is usually <u>not</u> known gives rise to the so-called *"false precision problem"*, by which is meant that values of which the accuracy is not known are treated as being exactly correct. All this means that a careful validation of a target acquisition model should be carried out before conclusions can be drawn from its predictions.

Model validation can be done by either analyzing the structure of the model, or by comparing the model's predictions to the actual performance of observers in military (field) tasks. This study takes the latter approach, by first measuring observer performance on a large number of thermal images collected during the field trial BEST TWO (organized by NATO AC243/Panel4/RSG.15, 1990), and then calculating the corresponding TARGAC predictions.

2. OBSERVER EXPERIMENTS AND FIELD TRIALS

Observer experiments must be carried out using carefully controlled procedures, and according to a design that allows proper statistical analysis of the data. For this reason, such experiments are best carried out in the laboratory, where conditions can be controlled and repetition of identical experiments with different observers is possible. In a field situation, conditions are difficult to control and often change quickly. Further, it is difficult to obtain accurate and reliable data on observer performance.

The field trial BEST TWO provided an opportunity to collect imagery for laboratory observer performance experiments that would yield data with sufficient accuracy for a quantitative evaluation of target acquisitie models. During the test, a large amount of thermal images of stationary and moving target vehicles at many distances were recorded. For these images target recognition and identification performance were determined. A limited observer experiment was carried out in the field for validation of the observer scores that were measured in the laboratory. Also collected during the trial were meteorological data, target contrast values and other parameters that are required by TARGAC to make acquisition range predictions.

3. OVERVIEW

The work carried out within this project consists of three studies. The experimental methods used are described in Chapters 2 and 3, and the results of the three studies in Chapters 4, 5 and 6. These chapters are based on a number of mostly confidential reports that were published earlier, see the bibliography. Most of the data in the original reports are confidential because they reveal recognition and identification performance for thermal imaging of a number of actual military targets. In order to make the present report unclassified, the targets have been coded with the letters A to I in all instances where data and graphs are concerned. This does not in any way influence the conclusions that can be drawn from this study. The key to the target codes is available on request, from either TNO-TM or ARL-BED (Dr. P. Gillespie). In addition to coding the target vehicle names, the contents of the original reports have been slightly rearranged, in order to obtain a coherent final report. Each chapter can, however, still be read more or less independently.

3.1 Field test

In Chapter 2, an overview of the NATO BEST TWO field trial is presented, and the scenario's carried out in the field and the different recording conditions are described. The weather during the test was hot and dry, and very constant over the three week test period.

This means that the effect of the weather on target acquisition performance could not really be studied.

3.2 Design of laboratory experiments

Chapter 3 treats the design of the laboratory observer experiments, and the training and selection of the observers. It is shown that:

- For the restricted set of target vehicles that was used, observers without any experience could be trained to an acceptable level in a few hours.
- Large differences in performance between subjects occurred, and within a few hours good observers could be distinguished from poor ones. A few observers were exceptionally good, and about 30% of the observers were unsuitable for the task.
- Both military and civilian observers were tested and no overall difference in performance between the two groups was found. In Chapter 5, however, it is shown that there is a difference in response *behaviour*.

3.3 Study I: Observer target acquisition performance

In Chapter 4 the results of Study I are presented. Target recognition and identification probability was determined as a function of range, for a number of different conditions. The effect of target type, time of day, approach route and target motion on the observer scores was determined. Further, two different ways of presenting the targets were used: 'pop-up', and 'sequential'. In the 'pop-up' presentation mode, randomly picked targets appeared at random positions in the field; in the 'sequential' mode, the targets were presented as an ordered sequence of decreasing distance, from 4 km on down. The latter condition simulated a target approach, during which the observer may accumulate information on the target. Both ways of target presentation have military significance. Search and target detection were not studied.

The data were collected in 2 main experiments in which a total of 24 observers were used. A total of 811 different images containing single targets were presented, and each target presentation was repeated 5 times (at random) in the course of an experiment. All data points in the graphs are the averaged scores over all participating observers. A complete overview of all the observer performance data that was collected is presented in Appendix 1.

The results can be summarized as follows.

- Identification and recognition performance varies considerably for the different targets in different trials and under different conditions: in some cases targets were recognized at 4 km while in others they could not be recognized at 1 km.

- Identification and recognition performance for a target that suddenly appears at a certain location ('pop-up') may be considerably worse than for a target seen approaching from a large distance ('sequential'). Current target acquisition models do not take this difference into account.
- Head-on target motion, either with or without a (dark) dust cloud behind the target, does not have a large influence on identification or recognition performance. It may, however, have an effect on detection probability.

3.4 Study II: Reliability of observer responses

Chapter 5 describes Study II, in which the reliability of observer responses and some of the more psychological and task-structure effects on the observer performance are analyzed. In a target acquisition task, there is always a possibility that an observer makes a wrong judgement. This can make the observer feel unsure and make him hesitant to give a report, although he may have recognized a target correctly. On the other hand, if he is confident enough to give a report or to undertake an action, it is of obvious importance to know the probability that he may be wrong, and which factors influence this probability. Therefore, apart from the observer's *skill* to identify or recognize a target, observer *confidence* (and the corresponding response *behaviour*) forms an important factor in target acquisition performance. This study was designed in such a way that the influence of *skill* and *behaviour* on target acquisition performance could be analyzed separately. Also, the reliability of first reports (that is the first time that an observer reports an identification or recognition) during a target approach was analyzed. The results show that:

- Observers that have the same skill can differ widely in behaviour. In practise this means that one observer will give his reports at much closer target ranges than another, although in principle they could provide the same information at the same distance.
- There is a large range of target distances at which the observers do not feel sure enough to report an identification, but respond correctly if they are forced to. Thus, if the circumstances ask for it, acquisition range may be significantly increased by forcing the observers to respond. Since this may also lead to an increase of false alarms, the instructions should be given depending on the circumstances.
- The reliability (the probability of being correct) of a first identification or recognition report during a target approach, is 80%, averaged over all subjects. This percentage is largely independent of target distance, target type, part of day and target background. Large individual differences (between 55% and 97%) are however found.

3.5 Study III: TARGAC validation

In Chapter 6, finally, the results of Study III are presented. A comparison is made between TARGAC *recognition* predictions and measured observer performance for a large number of trials.
First, TARGAC was subjected to a sensitivity analysis, in which the effect of changes in the input parameters on the model output was determined. This showed that, for the BEST TWO situation, the predicted recognition ranges are almost independent of time of day, and target- and background temperatures. Only the target effective dimension and the thermal imager resolution limit (the cut-off frequency of the Minimum Resolvable Temperature Difference curve, MRTD) had a significant effect on model output. This result was in fact caused by the excellent atmospheric conditions during the BEST TWO field test. It means that, in this study, the recognition performance predictions are determined solely by the system performance module of TARGAC, which is equivalent to the 1-dimensional NVESD Static Performance Model.

The model was validated by determining the ratio between measured and predicted acquisition ranges for a large number of trials, i.e. a large number of target approaches in different conditions. The results are expressed as a probability distribution of this ratio. The *mean* of this distribution quantitatively shows how well the model predicts *overall* acquisition performance over a large number of trials. The *variance* of the distribution is a quantitative measure of the accuracy of the model predictions for *individual* trials. If the mean of the distribution were equal to 1 and the variance equal to 0, then the model predictions would be perfect. Deviations of these ideal values indicate the extent to which to the model can be used. The results for TARGAC in the BEST TWO situation are as follows:

The mean of the ratio distribution is 1.8 (+/- 0.2). This means that, on average, observer performance is considerably better than the model predicts, and that a correction factor of 1.8 should be used to match the predicted recognition ranges to the results of the observer experiments. The *shape* of the mean probability vs range curve, however, is very similar for the model and the observations.

The *variance* of the ratio distribution is large: the ratio between measured and predicted acquisition range spreads between 0.9 and 3.6 (95% level), i.e. it spans a range of a factor of 4. Thus, TARGAC predictions of recognition range for individual targets can differ from the actual recognition range by a factor between 0.9 and 3.6, which means in fact that TARGAC does not predict recognition performance very well. A similar conclusion may hold for other models that are based on the same principle.

The TARGAC output was compared to the 1- dimensional ACQUIRE model, which showed that both models produce identical recognition vs ranges curves, i.e. both models under-predict the mean recognition range by a factor of 1.8. Predictions of the newer 2-D

ACQUIRE model, which uses the Johnson criteria in 2-dimensions, yielded a much better result: the factor of 1.8 was reduced to about 1.3. The variance in ratio distribution, however, remained the same when the 2-D model was used.

Software quality
The sensitivity analysis and a number of tests that were performed on TARGAC beiore the actual validation was carried out, brought to light a number of problems and errors in the software. These were fixed after consultation with Dr. Patti Gillespie from ARL-BED before the validation was carried out. Most of the problems have now been taken care of, and the program manual has been adapted.

4. CONCLUSIONS AND RECOMMENDATIONS

The target acquisition model TARGAC was validated using BEST TWO observer performance data for recognition of targets in front view. The major conclusions of studies I and II are, that human acquisition performance depends considerably on factors such as target structure, local terrain structure, and cognitive factors.
In TARGAC, and in many other target acquisition models, these factors are not incorporated, and study III shows that the TARGAC predictions for the BEST TWO situation hardly depend on the experimental and field conditions. Therefore, important differences between measured and predicted recognition performance were found.
The main results of the validation are:

1. The model predictions are too conservative. On average, TARGAC underestimates recognition range by a factor 1.8 (+/- 0.2).
2. TARGAC does make accurate predictions for individual cases. The analysis shows that the uncertainty interval roughly ranges from 0.9 to 3.6 times the predicted acquisition range, thus spanning a range of a factor 4.
3. TARGAC recognition performance predictions for the BEST TWO situation (excellent weather), are determined solely by its system performance module, which is equivalent to the 1-dimensional NVESD Static Performance Model.
4. The TARGAC predictions for *overall mean* performance can be improved by incorporating the 2-dimensional version of the Static Performance Model. For *individual cases*,, however, the predictions with the 2-D version are not better than those with the 1-dimensional version.
5. It is proposed that TARGAC predictions are not only presented as single numbers for acquisition probability *vs.* target range, but that some indication is given of the accuracy of the results, preferably in the form of a 95% confidence interval.
6. The version of TARGAC that was tested (PC version, released in June 1992) contained a number of software errors and some minor problems. A number of corrections are suggested. Additional work in modularizing and streamlining the model is

recommended. It is also recommended that the model is given a more consistent and user-friendly user interface.

7. TARGAC, and other models that incorporate the NVESD Static Performance Model should only be used to provide an *indication* of the actual acquisition performance.

CHAPTER 2

THE BEST TWO FIELD TEST

SUMMARY

In July en August 1990 the field trials BEST TWO (Battle field Emissive Sources Test under european Theater Weather and Obscurants) were held in Mourmelon in France. The test was organized by NATO AC/243, Panel 4, RSG 15; the participating countries were: United States, England, Germany, France, Denmark and The Netherlands. The purpose of the test was to quantify the performance of electro-optical imaging and observation devices under battle field conditions. This report gives an overview of the four scenarios that were carried out.

In Scenario 1, single target vehicles (tanks, Armed Personnel Carriers and Trucks), drove down a predefined track, from a distance of 4000 m towards the Main Instrumentation Area (MIA). The targets stopped for 2 minutes at designated positions along the track, roughly every 300 m. Two different tracks were used, one with 16 stop positions, and one with 10 stops. This scenario allowed the recording of imagery of stationary targets at a range of distances between 4000 m and 1000 m. Scenario 2 was very similar to Scenario 1, the difference being that the target vehicles did not stop. Battlefield effects were included in some versions of Scenario 2.

In Scenario 3, a column of either 8 or 12 target vehicles drove along a track across the field of view. The purpose of this scenario was to study the recognition of groups of targets, and the effects of dust on observer/system capability. In Scenario 4 an attack formation of 4 Tanks and 8 APCs approached the MIA from a distance of 4 km. The purpose of this scenario was to assess the effects of simulated artillery barrages and/or smoke on observer/system performance in a realistic task environment.

1 INTRODUCTION

The BEST TWO trials were held at "Camp de Mourmelon", near Chalons-sur-Marne, in the Northeastern part of France, from 26 July to 10 August 1990. This paper describes the 4 scenarios that were carried out during BEST TWO. A detailed description of target vehicle order, timing and movements during the scenarios, as well as the timing of the simulated battle events is presented elsewhere (Valeton, Bijl & De Reus, 1992).

1.1 General information

Session organization
The field trial was organized in 4-hour sessions, and 2 or 3 sessions were carried out per twenty-four hour day. The session time slots that were used are listed in Table I. The sessions were numbered by a code xx.y, where xx is the day of the month and y is the slot number (see Table I). Since the experiments were carried out in the last week of July and the first 10 days of August, use of only the day numbers allows a unique session code: session 27.3 is the afternoon session on July 27, and 3.4 is the late night session on August 3.

Table I Session time slots.

slot number	time period
1	02:00 - 06:00 hour (early night)
2	09:00 - 13:00 hour (morning)
3	14:00 - 18:00 hour (afternoon)
4	22:00 - 02:00 hour (late night)

Terrain layout
The basic pattern of the experiments was that one or more target vehicles was either stationary or moving in the terrain, while measurements and recordings were made from several sites in the field. Seen from the Main Instrumentation Area (MIA), the terrain was roughly 2 km wide and 4 km long. In most scenarios the target vehicles approached the MIA from the far end of the field up to 1 or 2 km from the MIA. Maps showing the target vehicle routes will be presented for each scenario in section 2. These maps were made with the field survey data provided by France. The target vehicle routes were marked in the field with numbered signs. The numbers made it possible to (repeatedly) position the targets at (the same) designated points, as for example in Scenario 1. At night the signs were dimly illuminated to allow the target vehicle drivers to find their way. Different color signs were used for the routes for the different scenarios: white and black for Scenario 1 and 2 (left and right), yellow for Scenario 3, and blue, blue/black and

black/blue for Scenario 4. Sometimes the approach routes are named after the color of the signs used.

Test time
Standard test time was provided by France and distributed in IRIG-B format to all national setups in the MIA on a coax cable. This signal was to be recorded with the measurements to provide a common time base for all data. The code was for example recorded on the audio channel of the video systems that were used to record imagery.

Laser experiments
Lasers were used during a large number of experiments, and during about half the sessions all personnel in the field was required to wear laser safety goggles.

Characterization
In addition to the four scenarios described in this report there were three "Characterization" sessions, in which physical measurements of isolated battle effects were made. One or two tanks were used as targets during these measurements, but no large scale vehicle movements were involved.

1.2 Target vehicles

A total of 14 target vehicles, of three different types were used. Two types of tank, 3 types of Armored Personnel Carrier (APC), and 1 type of truck. The French supplied 3 AMX-30 tanks, 6 AMX-10 APCs and 2 trucks. The Dutch contributed a Leopard 2 tank and two versions of the YPR 765 APC: the YPR-PRI, a standard APC with a 25 mm cannon, and the YPR-PRAT, which has a dual TOW anti tank missile turret. Germany added special thermal camouflage materials to three of the French vehicles, and these are identified with a "C" behind their name. Table II sums up the target vehicle that were used.

Table II Target vehicles.

Tanks	APCs	Wheeled
Leopard 2 AMX-30 AMX-30 C	PRI PRAT AMX-10 AMX-10 C	Truck Truck C

In order to have an "operational signature" during the test, the drivers exercised their vehicles for 15 - 30 min prior to each run, outside of the view from the Instrumentation Areas. There was a "signature post" at the back of the field, where the Danish and the Germans took calibrated measurements of all target vehicles before each run. If the signature was not right, the driver had to go back to further warm up his vehicle.
Target vehicle management was done by a cadet of the Dutch Royal Military Academy, who was in radio contact with the general test management in the MIA.

1.3 **Battle events**

During the test a large number of simulated battlefield effects were produced:

a) "sandbags" - a simulation of an artillery barrage by exploding sandbags that were suspended from tripods, each sandbag representing two 122 mm and one 152 mm Soviet shells.

b) "LUST" - canisters producing white phosphorous smoke.

c) "fires" - burning fuel and tires in oil drums as a simulation of a burning object on the battlefield.

Details of the execution of the simulated battle events are described in Danielian (1992). The scenarios 2, 3 and 4 were all carried out a number of times, with and without several of these simulated battle events.

2 SCENARIOS

2.1 **Scenario 1**

In Scenario 1 single target vehicles drove down one of 2 different approach routes ("left" and "right"). The routes are shown on the map in Fig. 1, which also shows the location of the MIA. The numbers along the tracks are the locations of the numbered signs. The distances between the stop positions and the MIA are given in Valeton, Bijl and De Reus (1992). In this scenario, the target vehicles stopped for about two minutes at each stop sign, allowing the teams in the instrumentation areas to make recordings and measurements of *stationary* targets. A stationary target in this case is not only a target that does not move, but also a target that has no large dust cloud behind it. The left route came as close as 1000 m from the MIA, while the closest distance for the right route was about 1600 m. Note that stop sign # 14 is missing on the left route; it disappeared during the first day of the test. Scenario 1 was carried out 6 times; 3 times along the left route, and 3 times along the right route. No battlefield effects were included.
Several nations had observers in the MIA doing a real-time target acquisition task. The order of the targets in each sessions was chosen at random (by the test management) so

the observers never knew in advance which target they were going see. A session lasted 4 hours, each target vehicle run lasted about 30 min, and on average 6 - 8 targets could participate in each session.

2.2 Scenario 2

Scenario 2 was very similar to Scenario 1, the only difference being that the target vehicles drove down the approach routes continuously, without stopping. The location of each vehicle during a run was later determined, see Appendix 1. This scenario allowed recordings and measurements of *moving* targets. Scenario was carried out 6 times, under different conditions. Two target vehicle speeds were used: 20 km/h ("fast") and 8 km/h ("slow"). The purpose of these two speeds was to create one condition of moving targets with dust thrown up by the vehicles, and one without dust.

Table III Details of Scenario 2.

scenario	route/speed	session #	battle effects
2A	left/slow	31.4	none
2A	left/fast	31.3	none
2A	right/slow	8.1	none
2B	left/slow	3.4	6 fires
2B	left/fast	1.3	6 fires
2C	left/slow	31.2	2 sandbags at 3500 m 2 sandbags at 2500 m

In practice it appeared that there was little difference in the amount of dust produced by the targets in both conditions. Two kinds of battle field effects were used: fires and artillery dust simulations. Table III gives the conditions used for the 6 versions of Scenario 2.

In Scenario 2C, one sandbag was exploded 50 m to the left of the route and one 50 m to the right of the route, both at 3500 m and at 2500 m, during each target vehicle run. The timing was such that the explosions went off when the target was right between the two bags.

In this scenario, the targets could complete a run in less, sometimes much less, than 30 min. However, in order to coordinate target movements with battle events and the timing of helicopter and fixed wing aircraft overflights, the targets were scheduled to start at exactly the half hours on the clock. This meant that usually only 7 runs could be completed and hence that not all targets could be included in each session.

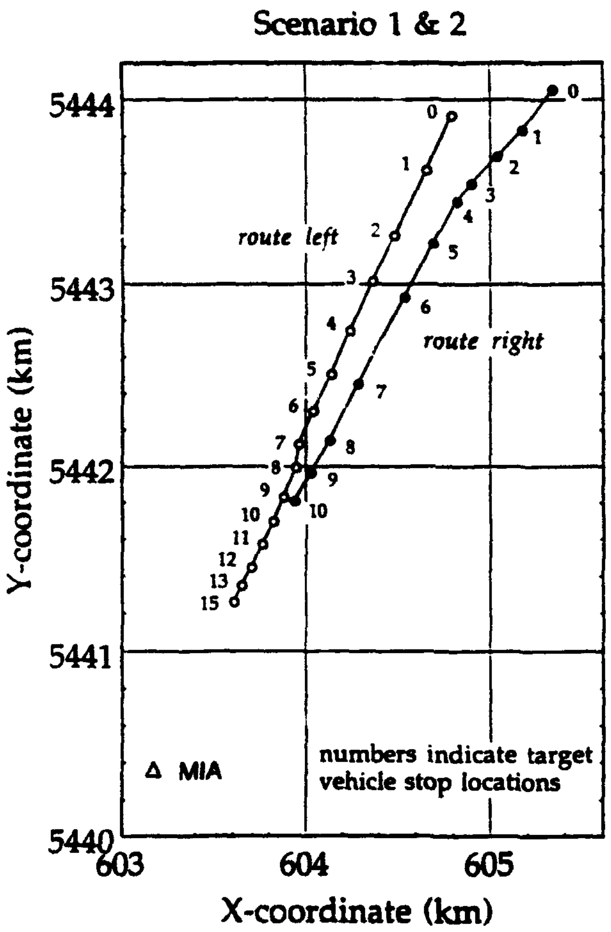


Fig. 1 Approach routes for Scenario 1 and 2.

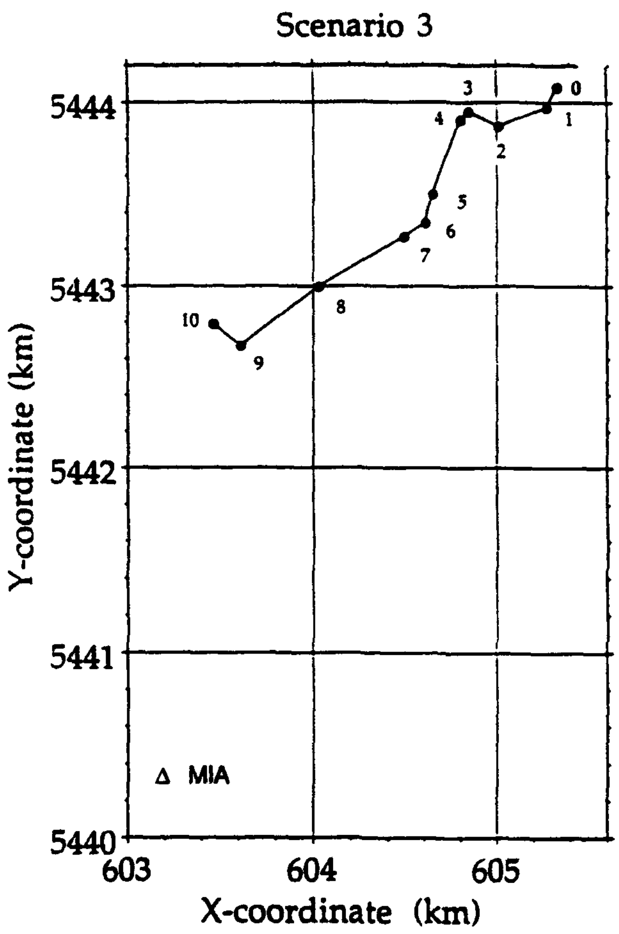


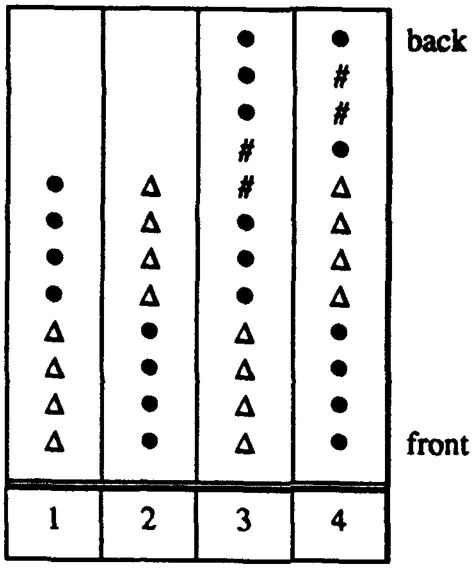
Fig. 2 Approach route for Scenario 3.

2.3 Scenario 3

In Scenario 3, a column of target vehicles, mutual distance 50 m, drove down an oblique track across the field, see the map in Fig. 2. Vehicle speed was 20 km/h.
The size and composition of a column of vehicles is important information. When a column is partly obscured by dust or battle effects, correct appraisal of the situation might be hampered. The purpose of this scenario was to determine whether observers can classify a column on the basis of its whole appearance, or "gestalt", instead of on seeing all individual vehicles.

To test this idea, four main types of columns were used. All columns consisted of Tanks and APC's, and the Tanks were always grouped together. The columns were either "short" (8 targets) or "long" (12 targets. In the short columns the Tank-group was either in front or at the back; in the long columns it was in front or in the middle. The four different formations are illustrated in Fig. 3. The short column was typical of two platoons moving to contact; the long column was typical of a company-size column of mixed composition.

The distribution of the available tanks, APCs and trucks within the layout of a given column was different for each run.



1	2	3	4	
		●	●	back
		●	#	
		●	#	
		#	●	
●	Δ	#	Δ	
●	Δ	●	Δ	
●	Δ	●	Δ	
●	Δ	●	Δ	
Δ	●	Δ	●	
Δ	●	Δ	●	
Δ	●	Δ	●	
Δ	●	Δ	●	front

Δ = tank ● = APC # = Truck

Fig. 3 General composition of the four different column types of Scenario 3. The layout is presented as the targets were seen coming towards the MIA: the bottom ones are in front.

Scenario 3 was carried out three times, twice without battlefield effects and one with 3 fires located at 3500 m from the MIA along the track. Details are presented in the Table IV. In each session, 4 runs could be completed, so a total of 12 runs were realized during the whole test.

A number of observers from different nations made real-time observations for Scenario 3. The task of the observers was to first quickly determine whether a column was long or short, second to classify the type (Tanks in front or not), and third to name the targets in the column.

Table IV Details of Scenario 3.

scenario	session #	battle effects
3A	7.3	none
3A	8.2	none
3B	6.3	3 fires at 3500 m

The main conclusion from the field observations is that the columns, in this test, are never recognized at a glance; the idea that a column with a certain mission has a certain "gestalt" does not appear to be true. The observers judge each vehicle of a column as it comes into view, and they count them one by one in order to find out what kind of column they are dealing with. The columns throw up a lot of dust, and depending on the wind, a whole column may be obscured by the dust cloud generated by the first vehicle.

2.4 Scenario 4

General

In Scenario 4, an attack formation consisting of four Tanks followed by eight Armed Personnel Carriers approached the MIA from the far end of the field to about 2500 m. Speed was 15 km/h. The purpose of this scenario was threefold. First, to collect data on the number and duration of "holes" or visibility windows in dust and/or smoke clouds. Second, to determine how many targets of an attack formation can be seen at any one time. These two points are important for determining the effectiveness of a) guided missiles like the TOW and b) direct fire systems like tank and APC guns.

Table V Details of Scenario 4.

scenario	session #	battle effects
4A	27.2-30.3	None
4B	6.2	"Rolling Artillery Barrage" near targets, see text
4C	1.2-7.2-10.3	Artillery barrage in front of MIA
4D	2.3-9.2	WP smoke in front of MIA
4E	3.2-9.3	Artillery barrage + WP smoke in front of MIA
4X	9.3	None

the third purpose was to record realistic thermal and visual imagery of an attacking formation under various battle field circumstances, for training purposes. This scenario was carried out 10 times, in 5 different forms, and usually several runs could be made in one session. The different versions and the combinations of battlefield effects used are listed in Table V.

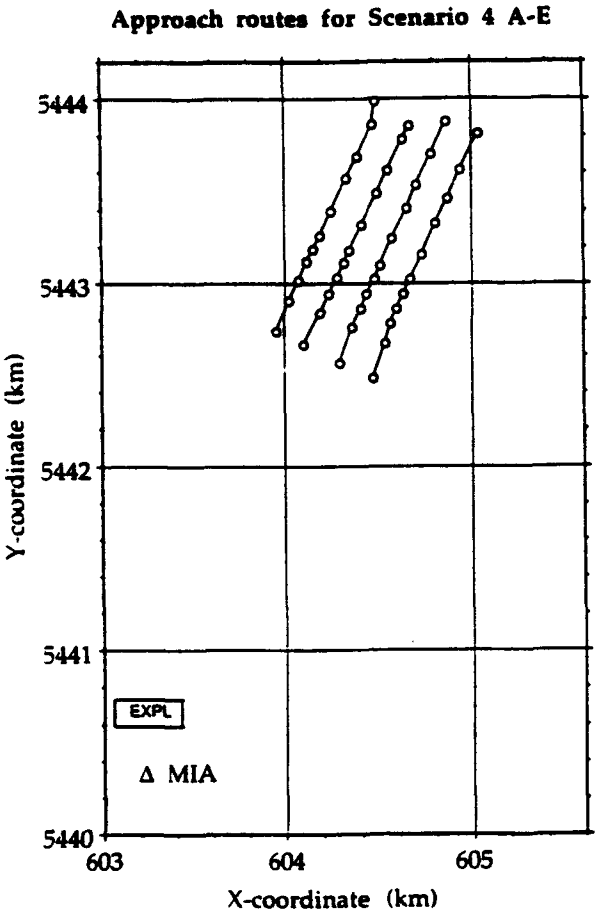


Fig. 4a Approach routes for Scenario 4A, C, D and E. The explosion area is marked "EXPL". Mutual distance between the tracks is 200 m.

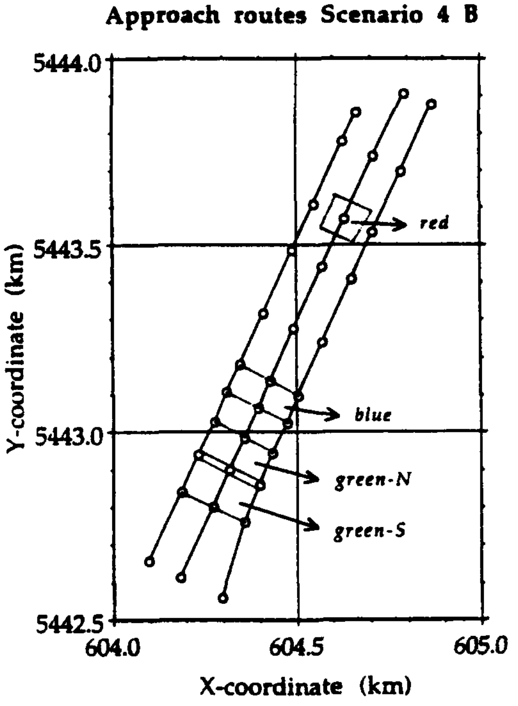


Fig. 4b Approach routes for Scenario 4B. The 4 explosion areas are indicated as: red, blue, green-N and green-S. For explanation, see text. The scale is twice that of Fig. 4a.

Scenario 4A was a baseline condition for Scenarios 4C, D and E. Scenario 4B was different from the others and will be described separately. The target tracks and explosion areas are presented in Fig. 4a and b. Scenario 4X was a short search experiment, and was different from all other versions of Scenario 4.

Scenario 4 A, C, D and E

In Scenario 4A, C, D and E, the attack formation was about 600 m wide. Four tanks were driving down the 4 approach routes shown in Fig. 4a, and 2 APCs followed each tank at a distance of 100 - 200 m. One of each pair of APCs drove 50 m left of the tank-track, the other 50 m to the right. The formation is schematically depicted in Fig. 5.

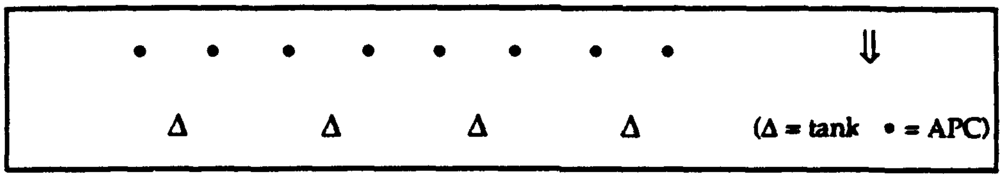


Fig. 5 Schematic diagram of the target formations for all versions of Scenario 4. The arrow on the right indicates the direction of motion.

The battle effects were produced in a 400 x 100 m explosion area, marked EXPL on the map in Fig. 4a, in front of the MIA, at a distance of 300 m.

Exp No. 1.2			Scen: 4C			
Rel	Abs					
Time	Time	EXPLOSION CREW			TARGET FORMATION	
(min)	(h:min)	Sandbags			(Vehicles use Blue/Black # signs)	
-50						Warm up of all vehicles
-10						Thermal signatures of 1 Tank & 1 APC
-5		Pre-check charges				Tanks to sign # 1; APC's to sign # 0
-4						
-3				Tank	APC	
-2				Pos. - Dist.	Pos. - Dist.	
-1						
0		Explode 1-st volley	(21 bags)	1 - 3700 m	0 - 3900 m	All vehicles WAIT
1		Explode 2-nd volley	(7 bags)	1 - 3700 m	0 - 3900 m	All vehicles WAIT
2		Explode 3-rd volley	(7 bags)	1 - 3700 m	0 - 3900 m	All vehicles WAIT
3		Explode 4-th volley	(7 bags)	1 - 3700 m	0 - 3900 m	All vehicles START; 15 km/h
4		Explode 5-th volley	(7 bags)	3 - 3400 m	1 - 3700 m	
5		Explode 6-th volley	(7 bags)	4 - 3150 m	3 - 3400 m	
6		Explode 7-th volley	(7 bags)	7 - 2850 m	5 - 3100 m	
7		Explode 8-th volley	(7 bags)	9 - 2650 m	7 - 2850 m	
8		Explode 9-th volley	(7 bags)	10 - 2400 m	9 - 2600 m	All vehicles WAIT
9						
10						
11						
12						All vehicles take tactical front-covered
13						positions facing MIA. Wait for 5 min.
14						
15						
16						
17		*** END SCENARIO ***				All vehicles leave field in single file,
18						followibg route Yellow.

Fig. 6 An example of a planning sheet: Scenario 4C.

These scenarios typically lasted 6 - 8 minutes. To illustrate the timing of target movements and battle events a sample of a planning sheet used for these scenarios is presented in Fig. 6. The scenario began after all targets were warmed up and had taken position at the beginning of the tracks. During the first 3 minutes, battle events (sandbag explosions, LUST devices, or both) were set off at 1 minute intervals in the explosion area in front of the MIA; the vehicles waited. After the dust/smoke screen was built up, the formation started to move and battle events were generated once a minute, for 6 more minutes, to sustain the obscuration. By the time the tanks reached the end of their tracks, at about 2500 m from the MIA, the event was over.

Scenario 4B

In Scenario 4B, a "rolling artillery barrage" was simulated. The idea here is that artillery paves the way for an attack formation by shelling (from behind) the area right in front of the own troops in order to neutralize all enemy defenses to enable fast and unobstructed advance of the own forces.

The attack formation had the same form as in Fig. 5. In the first phase (red) it was 200 m wide; after that it expanded to a width of 600 m. The tracks and explosion areas are shown in Fig. 4b, which gives an enlarged (2 x) view of the target area. The two outer tracks in Fig. 4b are the same as the two inner tracks of Fig. 4a. The central track in Fig. 4b was especially added for Scenario 4B and coincided with the left route for Scenario 1 and 2. The four tanks were spread out evenly over the 200 m width: one tank followed the leftmost track, one followed the rightmost track, one drove 30 m to the right of the central track and the other drove 30 m to the left.

Four explosion areas were laid out in the field and they were labelled "red", "blue", "green-N" (north), and "green-S" (south). In these four explosion areas different area/time densities of artillery bombardment were simulated, see Table VI. In area "red", for example, 20 sandbags, spread evenly over the 100x100 m, were exploded over a period of 2 minutes, while in both "green" areas the same number of bags was spread out over twice the area and exploded in a quarter of the time. These different conditions threw up dust clouds of different densities and duration.

Table VI Artillery barrage density in Scenario 4B.

Expl. Area	Size	# sandbags	Duration
red	100 x 100 m	20	2 min
blue	200 x 100 m	20	1 min
green-N	200 x 100 m	20	0.5 min
green-S	200 x 100 m	20	0.5 min

The execution of this scenario will now be described. The position of the formation is indicated as e.g. 3700/3900 m, where the numbers represent the distances of the tanks/APCs from the MIA.

Event RED

1 The target formation was stationary in position at 3700/3900 m.
2 The explosions in area red were set off; the targets did not move.

Event BLUE

1 The formation was moved 200 m down the tracks, to the blue start position at 3500/3700 m.
2 The targets started driving down the tracks at 15 km/h (= 250 m/min).
3 As soon as the formation was in motion, the explosions in area blue were started.
4 During the explosions the formation drove for about 300 m; they stopped when the explosions in area blue were finished.

Event GREEN

1 The formation was moved 300 m down the tracks, to the green start position at 3200/3400 m.
2 The targets started driving down the tracks at 15 km/h (= 250 m/min).
3 When the formation started moving, the explosions in area green were started: the first 1/2 minute in area green-N, the next 1/2 minute in area green-S.
4 During the explosions the formation drove for about 250 m; when the targets reached their end-point at 2900/3000 m the blue explosions were finished.

An analysis of Scenario 4 will be presented in Vonhof and Goessen (1992).

Scenario 4X

In this version of Scenario 4, the attack formation took position along ridge with some trees and bushes about 2000 m from the MIA. The vehicles were to take a position suitable for an attack on the MIA, but obscured from view by using the local terrain features as much as possible. The task of the observers in the MIA was to find the targets, on their thermal imagers. This scenario was carried out only once, just to get an idea of the detection probabilities in the described situation.

The (surprising) result of this short test was that only about half the targets were found, and that half of the responses that were given turned out to be false alarms (LCol. D.M. Vonhof, OCI, Royal Netherlands Army, personal communication).

3. TARGET POSITIONS IN SCENARIO 2

For the analysis of the data collected during the experiments the location and distance of each target must be known at all times. Since obtaining this time/distance information for Scenario 2 was not straightforward the procedure that was used is outlined below.

3.1 Radar data

The US fielded a multi-target tracking system (AUTOFEDS) to provide all target vehicle locations in x,y coordinates relative to the MIA as a function of time. Unfortunately this system was not operational during most of the test, and the limited data that came available proved to be not usable. Fortunately there were two French teams with a battlefield radar: 1) LMT-Radio Professionelle, from Boulonge and 2) RASIT, from ETCA, Arcueil. These systems were intended to provide backup information on target vehicle locations, and so information, albeit only for moving single targets, i.e. only for Scenario 2, was available after all.

In the analysis of the data provided by both radar systems two problems emerged. First, it appeared that in most of the data of the LMT radar, the compass bearing had not been set according to the local deviation of the earth's magnetic field. By rotating the coordinates of the vehicles by the amount of the local deviation, a reasonable fit with the land survey data could be obtained. This exercise proved that the LMT data were in principle usable. The RASIT radar produced geometrically correct data.
A second problem in the radar data concerns the time information. The master time for the test was provided by France and made available to all nations in the MIA in IRIG-B format, on a coax cable. We recorded the IRIG-B signal on the audio channel of our video recorders and hence had the correct time for the imagery on video. The computers of the radar systems had no hard-wired connection to the IRIG-B signal, and the computer clocks were presumably set by hand. This resulted, for many sessions, in differences between the time axes of the LMT radar, the RASIT radar and the IRIG-B time. We came across these problems when we compared time/distance information of the LMT and RASIT radars for a few runs with actual target time/position as inferred from video image/audio channel. Time differences between IRIG-B and the two radar systems of up to 2 minutes have been found. Since we believed we could not rely completely on the radar data, we determined the time/position relation of all targets in Scenario 2 by direct analysis of the video imagery.

3.2 "Visual" determination of target time/distance profile

The images of a complete Scenario 1 run were displayed on a monitor. A transparent sheet was attached to the face of the monitor, the approach route was sketched on the sheet and the locations where the target vehicles were stationary were marked. These marks correspond with the locations of the numbered signs along the approach route. This procedure was done for both the left and right approach routes. Next, imagery of the Scenario 2 runs was played back on the same monitor, while the IRIG-B time was displayed. At each instance when a target, as it moved along the track, passed a mark on the monitor face, the target was at a stop-sign location. The IRIG-B time of that instance was noted, and, since the distance of all numbered signs along the track are known, a series of correct time/distance values for each run was obtained. These data are indicated as "visual" or VIS. The VIS time/distance data were plotted, obvious errors were corrected, and comparisons with the radar data were made. A complete set of time/distance plots is presented in Valeton, Bijl and De Reus (1992).

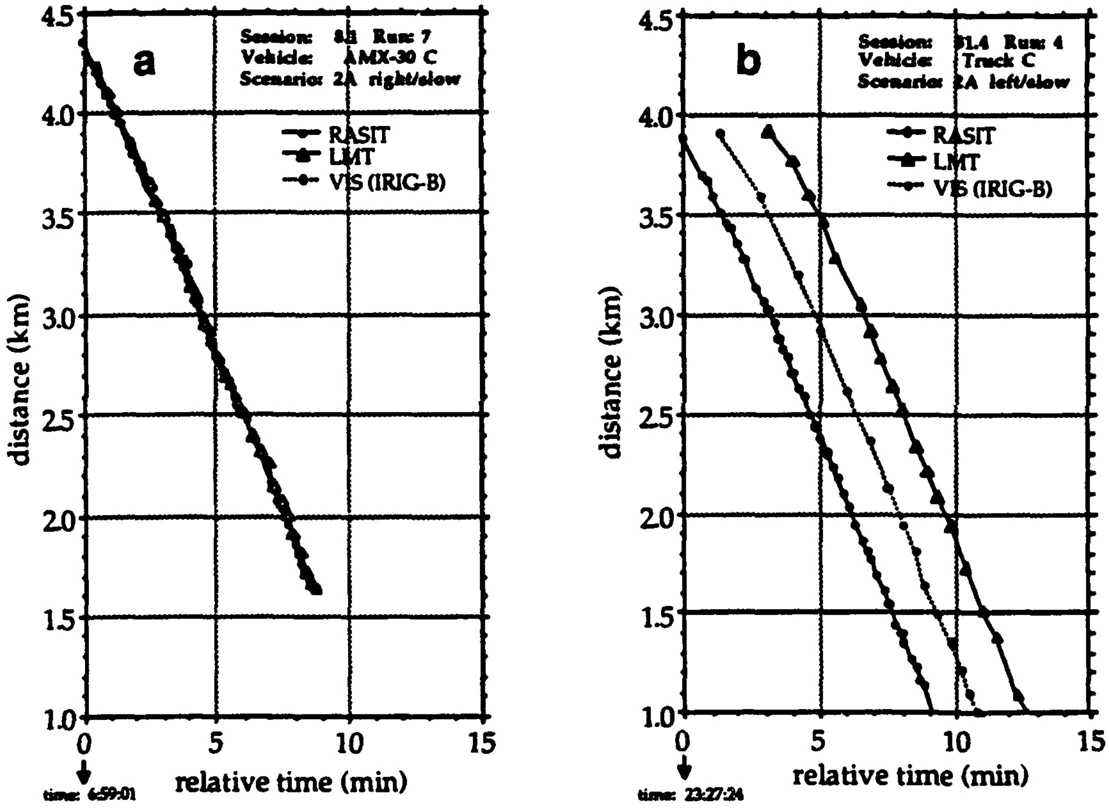


Fig. 7 Examples of time-distance relations for Scenario 2 from three different sources: the RASIT radar, the LMT radar and "visual" data obtained from IR imagery. In Fig. 7-a, left side, the data from the three sources coincide. In Fig 7-b, right side, time shifts between IRIG-B time and both radar data sets are apparent. These time shifts indicate errors in the time setting of the radar systems.

As an example of the differences in time/distance relations found for the three data sets, two extremes are presented in Fig. 7. In Fig. 7-a, on the left, a run is plotted for which the time/distance relations from all three sources coincide perfectly. Fig. 7-b, right side, shows an example in which there are time shifts of 1.5 and 2 minutes between IRIG-B and the radar data. Each symbol on a radar curve indicates a position generated by the radar. Note that the curves are almost straight lines. This shows that a) the target vehicle drove at constant speed, and b) that the three data sets are consistent and that the only difference is a shift in the time setting.

4 CLOSING REMARKS

The BEST TWO test was executed with great success. Only one of the planned sessions had to be cancelled, because the schedule was in fact overloaded. All national teams spent a great effort and collaborated successfully. In addition, France supplied field radio systems, the standard test time, a safety officer, road blocks, the explosives team that prepared, set up, and executed all battle event simulations (with a very high success rate). Further, a large number of logistic problems were solved every day. The weather during the test was mostly constant and very hot and dry. Meteorological data are presented in Smith and Corbin (1992).

In hindsight, it might be concluded that the test schedule had one flaw. Night sessions were to be included in BEST TWO, but it was decided not to sacrifice any daytime experiments. The night sessions were mostly planned so that there was a normal day with only 2 session between days that had night sessions. However, with this system, each late night session was followed by a normal 2-session day, which was again followed by a day with an early night session. This resulted in three consecutive days with very short breaks between sessions. Together with the large distance between test site and hotels and the extremely hot weather this made for a very strenuous test.

CHAPTER 3

DESIGN OF THE OBSERVER EXPERIMENTS

SUMMARY

In July and August 1990 the field trials BEST TWO (Battle field Emissive Sources Test under european Theater Weather and Obscurants) were held in Mourmelon in France. During the test, images of single stationary targets (Scenario 1) and moving targets (Scenario 2) were recorded with a thermal imager. The images are used in observer performance experiments to collect data for the evaluation of target acquisition models and for operational purposes. This report describes the methods and the design of the experiments. More extensive results of the observer experiments are presented in companion reports.

Stimuli consisting of short image sequences were shown to observers by using an analogue video disc system. This system allows presentation of events in random order at fast pace, while retaining the dynamic character of the imagery, for both stationary and moving targets. Identification and recognition performance was determined as a function of target range.

An extensive training program was developed; transfer of training to the main experiments was satisfactory. Large differences in performance between subjects occurred. The criteria for selection of observers for further analysis are discussed. Both military and civilian observers were tested and no overall difference in performance between the two groups was found.

1 INTRODUCTION

Thermal images recorded during the BEST TWO field trials (France, July-August 1990) were used in observer experiments in the laboratory to determine target acquisition performance. The purpose of these experiments is to collect data for the evaluation and development of target acquisition models, and to supply rules of thumb for operational purposes. The present observer performance experiments were restricted to target *identification* and *recognition*; Target *detection* and *search* are not studied. No battlefield effects were included.

The main results of the experiments will be described in the next three chapters; the present chapter treats the design of the experiments, the setup that was built, the observer training and the subsequent observer selection process. Also, a comparison is made between the performance of military and civilian observers.

Since detailed information on thermal image recognition on targets vehicles is confidential, target names are coded with the letters A-I in all data plots. This does not in any way affect the conclusions that can be drawn from the experiments see e.g. section 5.3. The key to the vehicle code is available on request.

2 METHODS

2.1 Field recordings

Four scenarios were carried out during BEST TWO. An overview of these scenarios has been presented in Chapter 2. Detailed information on the events during the trials (including test schedules, maps, time tables, vehicle positions and battlefield events) is reported in Valeton, Bijl and De Reus (1992). The present laboratory experiments were conducted with imagery collected during Scenarios 1 and 2. Recordings were made of stationary and moving single target vehicles at a range of distances between 4000 m and 1000 m. Nine target vehicles were used in these scenarios: three tanks, four APCs and two wheeled vehicles. Three vehicles were camouflaged.

The imagery was recorded from a 8-12 μm thermal imager on Umatic video tape. The field of view was 5 X 3 degrees (H x V). The camera was aimed such that the target vehicle was approximately in the middle of the image. A selection of the imagery was copied to video disc (see 2.2 and 2.3) for use in the laboratory experiments.

2.2 Experimental setup

A flexible setup was developed to present dynamic video imagery to four observers in parallel. The most important properties of the setup are described below.

2.2.1 *Dynamic stimulus display*

The heart of the setup was an analogue video disc system (Sony LVR-6000/ LVS-6000P) that was used to present the stimuli to the observers. This system is ideally suited for these kinds of observation experiments for two reasons. First, it allows the use of video sequences as stimuli, which comes as close as possible to the real field operation of thermal imagers because the image dynamics (spatio-temporal noise and image jitter) is retained. Both stationary and moving targets can be displayed realistically. Second, it allows the presentation of stimuli (short video sequences) to the observers in random order at fast pace. This is a requirement in the design of the observation experiments. The stimuli were displayed on Sony PVM 122 CE 12 inch monitors (white B4 phosphor) and the contrast and brightness controls were set for maximum linear contrast range before the experiment. The observers were not allowed to touch the controls.

The experiments were controlled by a PC that operated the video disc and was further interfaced (RS232) to four response panels that were used by the observers.

2.2.2 *Response panels*

Tandy Model 100 notebook computers were used as response panels. A number of keys on the keyboard were designated as target response keys by putting a name sticker on them. Six keys were assigned to the different targets: Leo, AMX-30, PRI, PRAT, AMX-10, and Truck. The camouflaged versions of the AMX-30, AMX-10 and Truck were not used as separate response categories. The reason is that we are interested in the ability of an observer to identify or recognize a target vehicle (either camouflaged or uncamouflaged), not in his ability to distinguish a camouflaged target from an uncamouflaged one.

Three keys were used to further qualify a response as an I(dentification), R(ecognition) or D(etection only). The latter responses are analyzed in chapter 5. The LCD display on the notebook computers was used to inform the observers on the status of the experiment and to give feedback during the training sessions.

2.2.3 *Observer setup*

The observers were placed in a dimly lit room and the response panel display was illuminated with a small external light source. Care was taken that no stray light fell on the monitor screen. The observers were allowed to choose their own optimal viewing distance and to scrutinize the display if they wished to do so. The viewing distance was 80 cm on average. The observers were watched from a control room with a closed circuit TV system.

An experiment lasted three days. The sessions lasted 30 - 45 min, and the observers were working in two shifts. While four of them were running the trials, the other four had a rest period.

2.3 Stimulus preparation

The video discs can store 24 minutes of video, or 36000 frames @ 50 Hz, per side. Because a large number of stimuli had to be stored on a single side, a limited number of frames was available per stimulus. Different sets of stimuli were used for the observer training and for the main experiments. In experiments where only stationary target images from Scenario 1 were used, the observers were trained with stimuli generated from Scenario 2 tapes (see below). In experiments where images from both Scenario 1 and 2 were used, about 40% of the available images were set aside for observer training.

2.3.1 *Stimuli for experiments*

For the images of stationary targets, sequences of 2 seconds (50 frames), taken at each vehicle stop position (see e.g. Chapter 2), were copied from the Scenario 1 video tapes to the disc. The required stimulus duration was however 5 seconds in the experiments and 4 - 9 s during training. These longer presentation times were obtained by playing the sequences repeatedly back and forth for as long as required. Smooth stimulus presentations of any desired duration could be obtained and it was not possible to see the difference with real continuous video, at least for stationary targets.

For the stimuli consisting of moving targets (Scenario 2), sequences of 5 seconds duration were copied to a separate video disc. In order to have the same observation distances as in the Scenario 1 imagery, the stimulus sequences were taken at exactly those times when the targets were passing the stop signs along the track. See Chapter 2 for details and maps of the approach routes for the two scenarios.

2.3.2 *Stimuli for observer training*

The observers were trained in four phases (see section 4). For the first two phases, a sequence of close-up images, recorded at 400-600 m distance during a characterization session were used to teach the observers the target names and their typical characteristics. Both thermal and visible CCD images of all targets, in front- and two side views, were copied to the video disc. In further training stages the images were similar to those used in the experiments. For experiments where only stationary targets were used, "stationary" training images were extracted from Scenario 2 (moving vehicles) by copying very short sequences of 15 frames to the video disc, and playing these back and forth. The moving targets now appeared stationary except at the shortest distances, where a slight rocking movement was visible. This made them a little easier to detect, which was considered an advantage in the training stage. Images of camouflaged vehicles were not used during training.

2.4 Stimulus presentation

Since in the present experiments we studied target *identification* and *recognition* only, the image presentations were structured such that the observers could always find the targets easily. Two ways of ordering the target images were used: "position" and "sequential".

2.4.1 *Position presentation*

In this case, targets "popping-up" at random positions were simulated by presenting targets at random distances along one of the two approach routes. First a position was chosen at random. In order to avoid search and detection problems, this position was shown to the observers. All target images available at that position (usually 6 - 8), were then presented in a randomly ordered sequence.

All images were presented to the observers five times (during different sessions).

2.4.2 *Sequential presentation*

A straight-line approach of the target towards the MIA was simulated by presenting the targets as an ordered sequence of decreasing distance, from 4 km on down. The reasoning behind this presentation method is that in a practical field situation targets may often be seen approaching for a certain period of time. During this time, as the target approaches, the observer may accumulate information on the target. This accumulation may possibly lead to a better acquisition performance than if the targets are presented at random positions. Sequential presentation of the images was repeated three times. The results of the comparison between the two presentation orders will be presented in Chapter 4.

2.5 Statistical analysis

The direct observer scores, i.e. the numbers of correct responses for all the conditions, are the data that were analyzed. If the observer chooses the correct target response key, he made a correct identification. If target and response belong to the same vehicle class (e.g. both are tanks or APCs), the response is a correct recognition. In this way, we obtained both identification and recognition scores in a single experiment. Most of the results in this and subsequent reports will be presented as plots of % correct responses (identification or recognition) *vs* target distance. In some of these plots error bars will be shown to give an indication of the accuracy.

The error in the observer scores can be calculated as follows. Let p be the probability that a target in a certain image will be recognized or identified correctly. Suppose that this image has been presented n times.

If the outcomes of the trials do not influence each other, this leads to a binomial distribution with mean value:

$$\text{score} \quad (\%) = p * 100\,\%$$

and standard deviation:

$$\sigma = \sqrt{\frac{p(1-p)}{n}} * 100\,\%$$

The standard deviation is highest if $p = 0.5$:

$$\sigma_{max}\,(\%) = \frac{50}{\sqrt{n}}$$

With $n = 100$ independent observations (e.g. responses from 100 different observers that see the image once), the maximum standard deviation in the data points will be 5%; with $n = 20$, $\sigma_{max} = 11\%$. In the experiment, each image was presented 5 times to about 20 observers. Since responses of the same observer cannot be regarded as independent, the maximum standard deviation falls between 5% and 11%. In some cases, less than 20 observers (but at least 5) were used. This leads to an increase in the standard deviation by a factor ≤ 2. The worst-case assumption therfore is: $10\% \leq \sigma_{max} \leq 20\%$. (Note that the actual values of σ will be much smaller when $p \neq 0.5$.)

Because the numbers of correct responses are distributed binomially, they were analyzed by using a log-linear model. This analysis is similar to Analysis Of Variance, which would be used for continuous, normally distributed variables.

2.6 Observers

Eight *civilian* observers, students from a nearby university, and seven *military* observers, (drafted) APC drivers/gunners, participated in the experiment. The military observers had a general military training, experience with a number of military vehicles, and with two kinds of thermal imaging systems. If no difference in performance between civilians and military personnel was found, future experiments can be carried out with paid volunteers, who are more readily available than military personnel.

All subjects were male and between 18 and 25 years old. They were tested for visual acuity before entering into the experiment. For all observers visual acuity was better than 1.5 arcmin^{-1}. Near vision acuity was tested with the TNO Priegel test; all observers scored better 20 mm^{-1}.

3 EXPERIMENTAL DESIGN

The present experiment was designed to measure the effects of different conditions (such as target presentation order (see 2.4), target motion, target camouflage, approach route, Part Of Day), or different groups of observers (see 2.6) on acquisition performance. Below, we will discuss some of the general design issues; some details are discussed more fully in the following chapters.

In order to draw statistically sound conclusions about performance in many different conditions, we need images of *all* targets vehicles in *all* conditions, at *all* ranges. This is called a *complete* design. With 9 target vehicles, 15 distances on the *left* route and 11 on the *right* route, and 4 different daily recording sessions (PODs), for both stationary and moving targets a complete design would consist of (9 x 15 x 4) + (9 x 11 x 4) = 936 images, that would have been collected in 72 runs. In practice it proved not to be possible to record all conditions: only 31 runs of Scenario 1 and 39 runs of Scenario 2 could be carried out. Four additional factors further decreased the number of available images: (1) the target vehicles were not, or barely, detectable at the first two stop positions of route Right, so 9 usable stops remained, (2) some vehicles went off track and got lost, (3) in Scenario 1, some stops on otherwise completed runs were missed, and (4) some images proved to be unusable because of unplanned disruptive events in the field. From Scenario 1, this left a set of 331 images for use in the experiments. The effective design for stationary targets is illustrated in Table I. Each dot represents a usable target image at a stop position. Missing dots in a sequence represent missed stop signs and the empty cells in the table indicate sessions that were not recorded at all.

In Scenario 2, the targets drove down the tracks without stopping, so in principle a very large number of short image sequences of moving targets can be extracted from the video tapes. Because image sequences from Scenario 2 were selected at the locations of the stop signs of Scenario 1, the complete design for Scenario 2 also has 936 stimuli, of which 480 usable images were available. The effective design for moving targets is similar to the one shown in Table I.

Table I Experimental design.

	Part Of Day 1		Part Of Day 2		Part Of Day 3		Part Of Day 4	
	Left	Right	Left	Right	Left	Right	Left	Right
Leo 2	••••• •••••••••		•••••••••••••••	•• ••••••		•••••••••		•• ••••••
AMX-30	••••••••		•••••••• ••••		•••••••• ••••••			•• •• •
AMX-30C	•••••••• ••••••			•••••••••	••••••• • •••			
PRI	•••••••• ••••••		••••• •• •••••	•••• •• •				
PRAT				•••• ••••	••••• •• ••••••	•••••••••		•••••••••
AMX-10			••••• •••••••••					•• ••••••
AMX-10C	•••••••••••••••				••••••• ••••••	•••••••••		•••••••••
Truck	•••••• •••••		••••• •••••••••	•••••••	••••••• •••••••	••••••••		••••••

Because the available image data is so incomplete, the effects of different conditions on acquisition performance had to be analyzed in smaller sub-designs. This means that a smaller number of vehicles (usually about 4) for which "balanced" information is available (e.g. targets that drove both approach routes at all possible PODs that can be compared) was used in the analysis. This reduced statistical significance of the results, but this problem is hard to avoid when imagery is collected in a large scale, multi-purpose field test.

4 OBSERVER TRAINING

4.1 Training structure

The observers were trained in 4 phases. The purpose was to bring the observers at about the same level of skill at the start of the experiments, and to avoid contamination of the data by the effects of possible additional learning during the experiments. The images used for training are described in section 2.3.2. Observer training took 4 - 5 hours of running trials and, with two groups working in shifts, was completed in one day.

Phase 1. The observers were shown sequences of 4 pairs of CCD and FLIR images of the same vehicle: 2 times a front view and both side views. The name of the vehicle was presented on the response panel display. Each image was shown for 4 seconds. The observers were encouraged to make notes and details were pointed out to them. The sequence was repeated four times. This phase lasted about 20 min.

Phase 2. The front view thermal images of Phase 1 were shown to the observers in random order, with a presentation time of 7 seconds. The observers responded by pressing a key corresponding to a target, and the response was echoed on the LCD display. Feedback was given by showing the correct target name on the display and by beeping when the response was wrong. This phase was repeated until all observers scored better than 95% correct. One session consisted of 3 presentations of all 6 vehicles and lasted about 4 minutes.

Phase 3. Images of all 6 target vehicles at distances of 1001, 1096, 1215 and 1350 m were used. The images were either taken from Scenario 1 or from Scenario 2, depending on the experiment. Each vehicle appeared 12 times, so a complete session consisted of 72 presentations. Presentation time was 7 seconds, and the response was followed by feedback. Phase 3 was repeated 4 or 5 times.

Phase 4. This Phase was very similar to Phase 3, but now closely resembled the real experiment. Target vehicle distances bracketed the complete range up to 4000 m, images were taken from both the Left and Right approach routes, and 4 PODs. A Phase 4 session consisted of 150 stimulus presentations: 2 x 8 distances x 6 vehicles from route *left* and

9 distances x 6 vehicles from route *right*. Stimulus presentation was 9 seconds in the first session. It was decreased to 7, and finally to 5 seconds in subsequent sessions. In addition to the target response keys, the "acquisition level" keys I(dentification), R(ecognition) and D(etection only) now also had to be used. The total duration of a session was 30 min and feedback was given. Phase 4 was carried out 4 or 5 times, depending on the scores.

4.2 Confusion matrices

A confusion matrix gives a quick impression of the results of an observer on a particular session. This matrix is a table that shows how the responses of the observer are distributed over the different targets. Two examples are presented in Fig. 1a and b.

a	Responses (%)					
	Tank		APC			Wheel
Test object	Leo 2	AMX-30	PRI	PRAT	AMX-10	Truck
Leo 2	**58**	8	17	17	.	.
AMX-30	17	**75**	.	.	8	.
PRI	8	.	**75**	17	.	.
PRAT	.	17	33	**50**	.	.
AMX-10	33	.	.	8	**58**	.
Truck	.	.	.	.	.	**100**

b	Responses (%)					
	Tank		APC			Wheel
Test object	Leo 2	AMX-30	PRI	PRAT	AMX-10	Truck
Leo 2	**100**	.	.	.	.	.
AMX-30	.	**100**	.	.	.	.
PRI	.	.	**83**	17	.	.
PRAT	.	.	8	**92**	.	.
AMX-10	.	.	.	.	**100**	.
Truck	.	.	.	.	.	**100**

Fig. 1 (a) Confusion matrix for an observer after only a few trai- ning sessions. Numbers are percentages of responses assigned to each response category. Correct responses are on the diagonal. Numbers in off-diagonal cells represent confusions between targets. The overall correct score is 69%. (b) Confusion matrix for the same subject after training was completed. Overall correct score: 96%.

The targets are listed in the left hand column and the response categories are listed in the top row. The row to the right of each target contains the percentages of responses in each category given to that target. Consider for example Fig. 1a. This is the confusion matrix for one observer after only a few training sessions. The first row of data shows that of all Leopard 2 presentations, 58% were correctly scored as Leo 2. However, the Leo 2 was seen as an AMX-30 in 8% of the presentations, and it was identified as PRI or PRAT in 17% of the cases. These errors are called confusions. The confusions for the other targets can be analyzed likewise. Note that responses on the diagonal (bold) are the correct responses. This observer, in this session, had a total percentage of 69% correct.

After training is completed, performance is much better, as can be seen in Fig. 1b. Here only a few confusion between PRI and PRAT remain.

Confusion matrices were calculated after each session and these were discussed with the observers to help them improve their training score.

5 RESULTS

5.1 Training results

The results of the training for the civilian and military observers, and an indication of their performance in the main experiment, are presented in Figs 2a and 2b. The data points present identification scores for individual observers, averaged over target type, approach route and Part Of Day. Three kinds of data are presented: a) the results of phase 3, session 1 - 5 (connected points), b) the average score for a subset of phase 4 trials, indicated as session 6, and c) the average score on a subset of the main experiment trials, indicated as session 7. Details on these subsets will be given below.

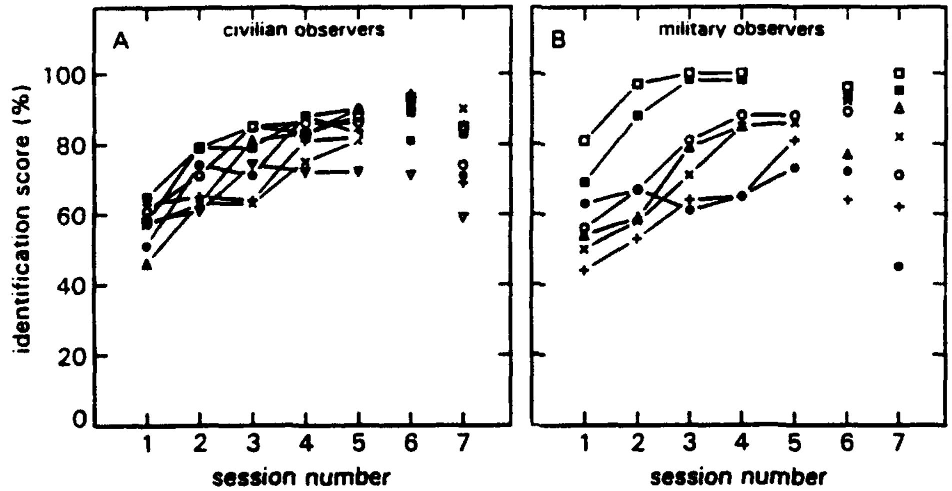


Fig. 2 Results of observer training, A: civilian observers; B military observers. Three data sets are shown in each panel: score as a function of session number for phase 3 (connected symbols, session 1 - 5), mean score for phase 4 (number 6) and mean score for the main experiment (number 7).

5.1.1 *Phase 3 results*

The general trend in the phase 3 results is that performance increases with session number, which means that the observers are learning. Most observers have reached a steady level after about 5 sessions, which indicates that no further learning occurs.

From Fig. 2a (civilians) we conclude that four observers are good: they have reached a high level at the final session of phase 3; three observers are considered reasonable and one observer is weak. Fig. 2b (military) shows that two observers are extremely good, three observers are reasonable and two are definitely weak. The civilian observers all show rather similar behavior, whereas there are large differences between the military observers. This difference can possibly be explained by the fact that the civilians, being students, were more homogeneous as a group and were used to sitting down at a task for long periods of time. The military observers (draftees), on the other hand, were not a homogeneous group with respect to educational level, and they exhibited large differences in attitude and motivation.

5.1.2 *Phase 4 results*

The second data set in Figs 2a and b (indicated as session 6) shows the average scores for the shortest target distances (those ranges that were used in phase 3) of the phase 4 trials. For almost all observers these scores are roughly at the end-level of phase 3. This means that there is sufficient transfer of training from phase 3 (only short ranges) to phase 4 (ranges up to 4 km).

5.1.3 *Main experiment means*

The third data set in Figs 2a and b (indicated as session 7) shows the average scores obtained for the shortest target distances (those ranges that were used in phase 3) in the main experiment. Most observers, except the weak ones, have a score that is only slightly lower than in the training. This means that there was sufficient transfer from training (with feedback) to the main experiment (no feedback). The slight drop in performance may be explained by the fact that different images were used in the main experiment, and that picture recognition (see below) was not possible.

5.1.4 *Picture recognition*

During the training process, some observers learned to recognize the *pictures* rather than the targets on the images. This was possible because of the feedback that was given after each target presentation. Since *target recognition* is the purpose of these experiments, *picture recognition* can in principle contaminate the data. The main reason to use different image sets for training and main experiment was to avoid the effects of picture recognition. Because no feedback is given in the main experiment, picture recognition cannot lead to higher scores.

5.1.5 *Observer selection*

The large differences in overall performance between observers brings up the question of observer selection. In psychophysical experiments it is in general not allowed to exclude observers that behave differently from the analysis because this obviously gives rise to biased conclusions. In the present case the situation is different. The final purpose of these experiments is to evaluate target acquisition models. These models should describe the performance of military observers in the field, and such personnel have usually gone through extensive training. It appears that not everybody can learn the task well enough to be an operator in the field: in practical military training about 30% of the trainees fail. For this reason we decided to set (in advance) a performance criterion and drop the worst observers from the final analysis.
The criterion is based on the following considerations: 1) at the shortest ranges (1000 - 1350 m) recognition is a relatively easy task; an observer that cannot distinguish a tank from an APC at those ranges with a reliability of at least 90%, is in fact not fit for the task. 2) During the BEST TWO field trials, real time observations by military observers yielded identification and recognition scores at short ranges of 75% and 95% correct, respectively. After consulting military experts, we decided that identification and recognition performance in the main experiments, averaged over all conditions and for short target ranges only, should exceed the following limits:

identification:	better than 70% correct
recognition:	better than 90% correct

These criteria resulted in dropping two civilian and two military subjects (which is about 30%).

5.2 Military *vs* civilian observers

Acquisition performance was determined for 15 runs. In Figs 3a, b, c and d, identification or recognition scores are plotted as a function of target range for four of these runs. The standard deviation (not plotted) in each data point is about 10-20%. Only the results of the "position" presentation (see 2.4) are shown. Solid lines represent performance for the civilians, dashed lines for the military observers. In all four examples the groups show a very similar, if not identical behavior. Second, the performance - distance relations are very different in the four examples. Since the scores for the two groups are so similar for such widely different conditions, we conclude that there is no difference in overall performance between well-trained military and civilian observers. Statistical analysis of the data shows no significant main effect on observer groups. This means that, in the further analysis, the results for all observers can be taken together in order to reduce the statistical errors.

5.3 General observations

The following general observations can be made. Target F on the right approach route, Fig. 3a, is identified correctly up to 3500 m. The combination of the thermal signature of the this target and the local background is apparently such that it stands out clear!y, almost to the end of the range. Identification of target A on the other hand, Fig. 3b, starts to become difficult already at ranges greater than 1500 m and is down to chance level at 2500 m. The general shape of this performance curve is what one would expect: a gradual decrease in score with increasing range, albeit somewhat steep in this case. A similar behavior is found in Fig. 3c.

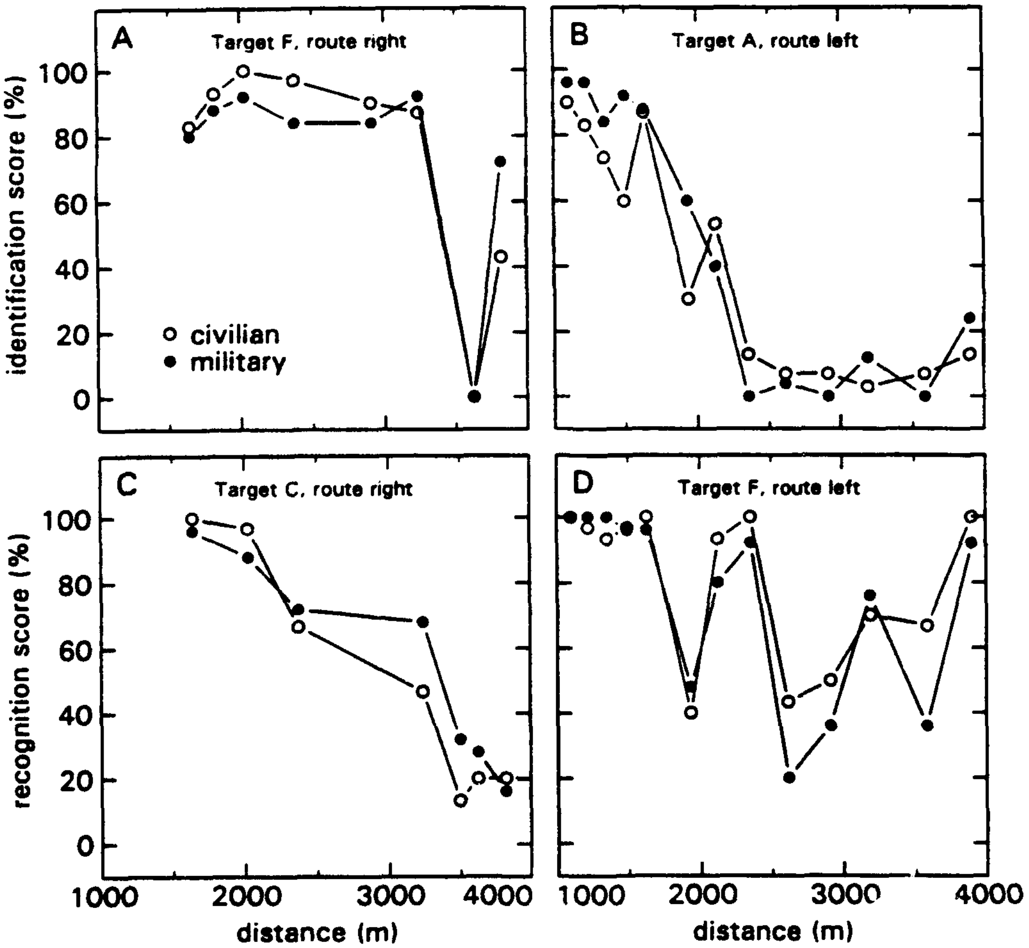


Fig. 3 Acquisition performance for civilian (open circles) and military (filled circles) observers. (a) Identification score for target F, route right; (b) Identification score for target A, route left; (c) Recognition score for target C, route right; (d) Recognition score for Target F, route left.

The curve in Fig. 3d (for a target F at the left approach route) is very different from the other ones but similar for both groups of observers. The local background on this left approach route at 2000 m, 2600 - 3000 m and at 3600 m is apparently such that this target is extremely difficult to identify or even recognize at those ranges, while at a range of about 2400 m or 3900 m it appears to be no problem at all.
A similar behavior was found for many different runs, and since it is caused by a strong interaction between target signature and local background, we term this behavior "target-terrain interaction". Chapter 4 will treat this phenomenon in more detail.

6 DISCUSSION AND CONCLUSIONS

The experimental setup that was built for these observer performance experiments proved to be a very practical and flexible tool. Training and experiments were set up quickly and were run very efficiently. Due to problems specific to large scale field trials, the design was incomplete from a statistical point of view. This problem was alleviated by using many observers and by repeating each experiment several times. This reduced the experimental error in individual data points to about 10%, which is sufficiently small for our purposes: the evaluation and development of target acquisition models, and the deduction of rules of thumb. Since target acquisition models in fact describe general behavior, we are interested in large effects, or the absence of effects, only.

The method of training the observers worked very well. Analysis of the performance of the military and civilians observers revealed no relevant differences. Note that this finding applies to this study only, where both groups went through exactly the same training procedure. We conclude a) that the observer training was sufficient, and b) that the task could be learned relatively easily.

CHAPTER 4

STUDY 1: OBSERVER TARGET AQUISITION PERFORMANCE

SUMMARY

In July and August 1990 the field trials BEST TWO (Battle field Emissive Sources Test under european Theater Weather and Obscurants) were held in Mourmelon in France. During the test, images of single stationary targets (Scenario 1) and moving targets (Scenario 2) were recorded with a thermal imager. The images are used in observer performance experiments to collect data for the evaluation and development of target acquisition models and for operational purposes.

The influence of head-on motion of targets, and the differences in acquisition performance for morning, afternoon and early night recordings were studied. In a number of situations, performance was heavily influenced by the interaction between the target and the local background properties ("target-terrain interaction"). Also, differences in performance for pop-up targets and approaching targets were found. Such differences are not described by current target acquisition models.

1 INTRODUCTION

Thermal images recorded during the BEST TWO field trials (France, July-August 1990) were used in observer experiments in the laboratory to determine target acquisition performance. The purpose of these experiments is to collect data for the evaluation and development of target acquisition models, and to supply rules of thumb for operational purposes. The design of the experiments, the setup that was built, the observer training and the subsequent observer selection process are described in Chapter 3; in the present chapter the main results will be presented, and the effects of several parameters on acquisition performance will be determined. A complete set of observer response data is presented in Appendix 1. The experiments were restricted to target *identification* and *recognition*; Target *detection* and *search* are not studied in the experiments.
The targets are indicated by the letters A-I because recognition performance data on these targets is confidential.

2 METHODS

The experimental method is explained in detail in ChApter 2. Briefly, a selection of the thermal imagery recorded during the field trial was shown to observers by using an analogue video disc system (Sony LVR-6000 / LVS-6000P). The experiments were controlled by a PC that operated the video disc and collected the responses from a maximum of four observers.
During BEST TWO, recordings were made of stationary (Scenario 1) and moving (Scenario 2) single target vehicles at a range of distances between 4000 m and 1000 m. About 10-15 short image sequences were taken from each run. In total, 24 scenario 1 and 14 scenario 2 runs were used in the laboratory experiments. The images contained one of the nine single target vehicles listed in Chapter 2, Table II. Three of these vehicles were camouflaged.
The images were presented to the observers for 5 seconds. The observer's task is to name the target, and to hit the designated key on a response panel after each presentation.
The analysis of the observer responses is explained in Chapter 3. The results will be presented as plots of % correct responses (identification or recognition) *vs* target distance. In some of these plots error bars will be shown to give an indication of the accuracy. See Chapter 3 for the calculation of these error bars. In most cases the standard deviation will be 10-20%.
Two different ways of ordering the target images were used: "position" and "sequential". In the *"position"* presentation, targets were presented at random distances along one of the two approach routes. In the *"sequential"* condition, the targets were presented as an ordered sequence of decreasing distance, from 4 km on down. During this time, as the target approaches, the observer may accumulate information on the target. This accumulation may possibly lead to a better acquisition performance than if the targets are presented at random positions.

The experiments were preceded by an extensive training. The purpose of this training is to bring the observers at about the same level of skill at the start of the experiments, and to avoid contamination of the data by the effects of possible learning during the experiments. Training material was selected from runs that were not used in the main experiment. The training method and results are discussed in detail in Chapter 3.

3 EXPERIMENTAL DESIGN

The main purpose of the experiments is to collect data for the evaluation and development of target acquisition models. For a thorough evaluation, it is important that acquisition performance is tested under a wide variety of conditions (i.e. for different times of the day, different atmospheric conditions, different terrain conditions, etc.). A second demand, especially important for model development, is to find out which parameters influence acquisition performance significantly. Therefore, it is necessary to vary one parameter, while keeping all other conditions unchanged. In order to draw statistically sound conclusions about performance in many different conditions, we need images of *all* target vehicles in *all* conditions. This is called a complete design. As was shown in Chapter 3, the available material is far from complete. This means that the effects of different parameters on acquisition performance have to be analyzed in smaller sub-designs, which reduces the statistical significance of the results. Other factors that complicate a comparison between different situations is that background temperature varied during the sessions, and that the field recordings were spread over several weeks. Luckily, the atmospherical conditions were remarkably constant during that period, but terrain conditions, e.g., changed significantly due to the intensive use of the approach routes.
We restrict ourselves to a few parameters that seem most relevant to target acquisition modelling. Those are: (1) target presentation order (position or sequential), (2) approach route, (3) target motion, (4) camouflage and (5) part of day. In detail, the following questions will be discussed in this chapter:

1 Does the presentation order influence acquisition performance? In other words, does accumulation of information lead to higher scores for an approaching target that has been spotted for a while, than for a target that suddenly appears at a certain distance?
2 Is observer performance different for the two approach routes (left, right), i.e. does local background structure influence performance significantly?
3 What is the effect of (head-on) target motion on observer performance, i.e. what is the difference in performance on Scenario 1 (no movement) and 2 (continuous movement) images?
4 What is the effect of the camouflage used on three of the vehicles on recognition and identification?
5 What is the effect of part of day on observer performance?

The data were obtained in two series of experimental sessions, with different groups of observers. The reason was, that not all questions could be answered in a single experiment. A second reason was, that about half of the image material had to be used for training purposes. By using different stimulus material and different groups of observers, a more complete dataset was collected.

In Experiment 1, acquisition performance was determined for 15 daytime runs (POD 2 and 3) of stationary targets. Both types of presentation order were applied. A group of seven military and eight civilian observers participated in this experiment. As was shown in Chapter 3, there is no difference in performance between these two groups. On the basis of a predefined criterion (see Chapter 3), 11 observers were selected. The mean scores of these observers will be presented.

In Experiment 2, data was collected for 33 runs of both stationary and moving targets on all PODs. Part of these runs were also used in Experiment 1. Only position presentation was applied. A group of seven (new) civilian observers participated in this experiment. After applying the criterion, only four observers were selected and their scores are used in the present report. However, two of the observers we dropped, performed only slightly below the selection criterion. We also calculated the mean results for six observers, and these are very similar to the results for the selected observers, except for a slightly lower overall score. Furthermore, the results for the identical runs in Experiments 1 and 2 are very similar. This means that, although the scores of only four observers are used in the analysis, the results are reliable.

4 RESULTS

4.1 Approaching and pop-up targets

In Figs 1a-f, identification scores, obtained in Experiment 1, are plotted as a function of target distance for six different runs. Open circles represent the scores for "position" presentation, filled circles represent the "sequential" condition. The runs are selected to illustrate three different types of acquisition performance vs. range we found for the position condition (open circles). In Figs 1a and 1b, identification score is invariantly high (except for one or two positions) and almost independent of target distance. Apparently, under these conditions target contrast and camera resolution allow excellent identification of the targets F and G, respectively, up to distances of at least 3500 m.

In Figs 1c and 1d, identification performance is good at short distances, but the score gradually decreases with target distance. For distances beyond 3000 m, the score is at about chance level (17%). Such a behavior is expected, e.g., if the resolution is too low to distinguish the various targets from each other at large distances. Another possibility is that meteorological conditions limit the identification range (which was obviously not the case during the BEST TWO trials). Also notice that the steepness of the curves in Figs 1c and 1d is different.

In Figs 1e and 1f, a very different relation between performance and target range is found. Performance depends largely on the exact position of the target. In these cases, camera resolution is not the limiting factor, because performance is quite good at the largest distance. Similar results were found for three observer groups.

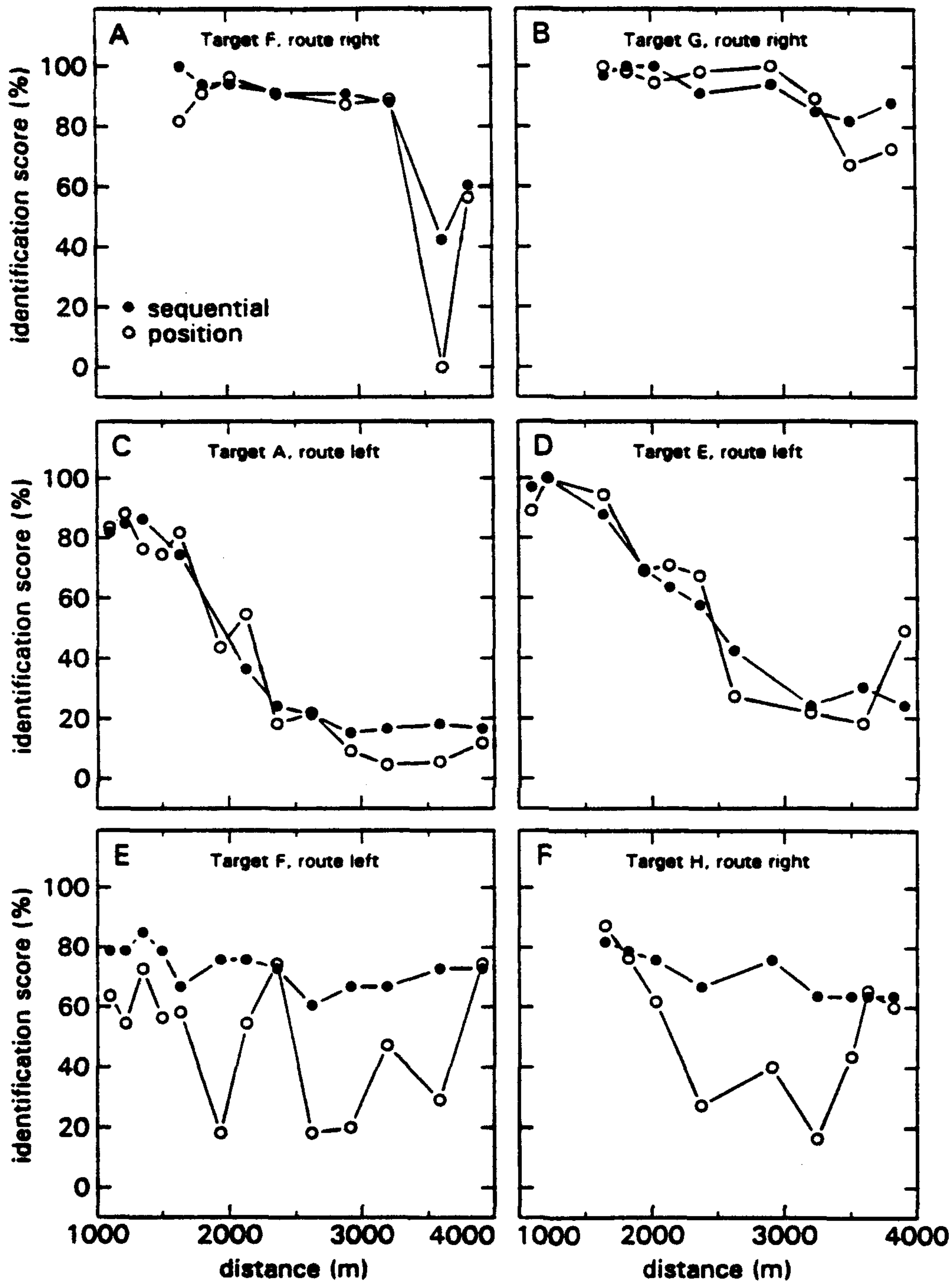


Fig. 1 Identification score as a function of target distance for six runs. Open circles: position presentation; filled circles: sequential presentation. a and b: Performance is invariantly good; c and d: Identification score decreases gradually with target distance; e and f: Large target terrain interactions are found for the position presentation. Sequential presentation leads to more stable results.

Apparently, in Fig. 1e the local background on the left approach route at 2000 m, 2600-3000 m and 3600 m is such that target F is extremely difficult to identify at those ranges, while at a range of 2400 m and 3900 m it is quite easy. Inspection of the imagery shows that indeed contrast between target and local background varies greatly during this run: in some cases identification is simple, in others the target is barely visible, and at a few positions it is easily confused with one of the other vehicles. A similar behavior was found for many other runs. Since it is caused by a strong interaction between target signature and local background, we term this behavior "target-terrain interaction".
The filled circles in Fig. 1 represent the scores obtained with the sequential presentation. In Figs 1a-d, the results for sequential and position presentation are similar. Apparently, information obtained from earlier presentations does not improve performance if the information from the actual presentation is equal or better. In Figs 1e and 1f, however, large differences between the curves appear. The scores obtained with the sequential presentation are much more stable with respect to target distance. At those positions where the information is sparse, the observers usually retain their choice from an earlier presentation. As a result rapid drops in performance, due to target-terrain interactions, do not take place during a target approach.

In the remainder of this chapter, only data obtained with the position presentation will be considered. These data contain the most complete information, since performance is based on single images without being influenced by history. This makes a comparison between conditions more sensitive to the local properties of target and terrain. On the other hand, if large differences between conditions are found, the differences may be less striking in the case of a target approach.

4.2 Part Of Day

In order to determine whether the observers perform differently on different times of the day, we analyzed the effect of the parameter "Part Of Day" on observer performance. There are only three conditions for which a comparison between more than two parts of day is possible. The recognition scores for these conditions are plotted as a function of target distance in Figs 2a-c. Although the amount of data is too limited to draw sound conclusions, Part of Day does not seem to have a large influence on observer performance. The variability in the data indicates that target terrain interaction plays a much more dominant role.

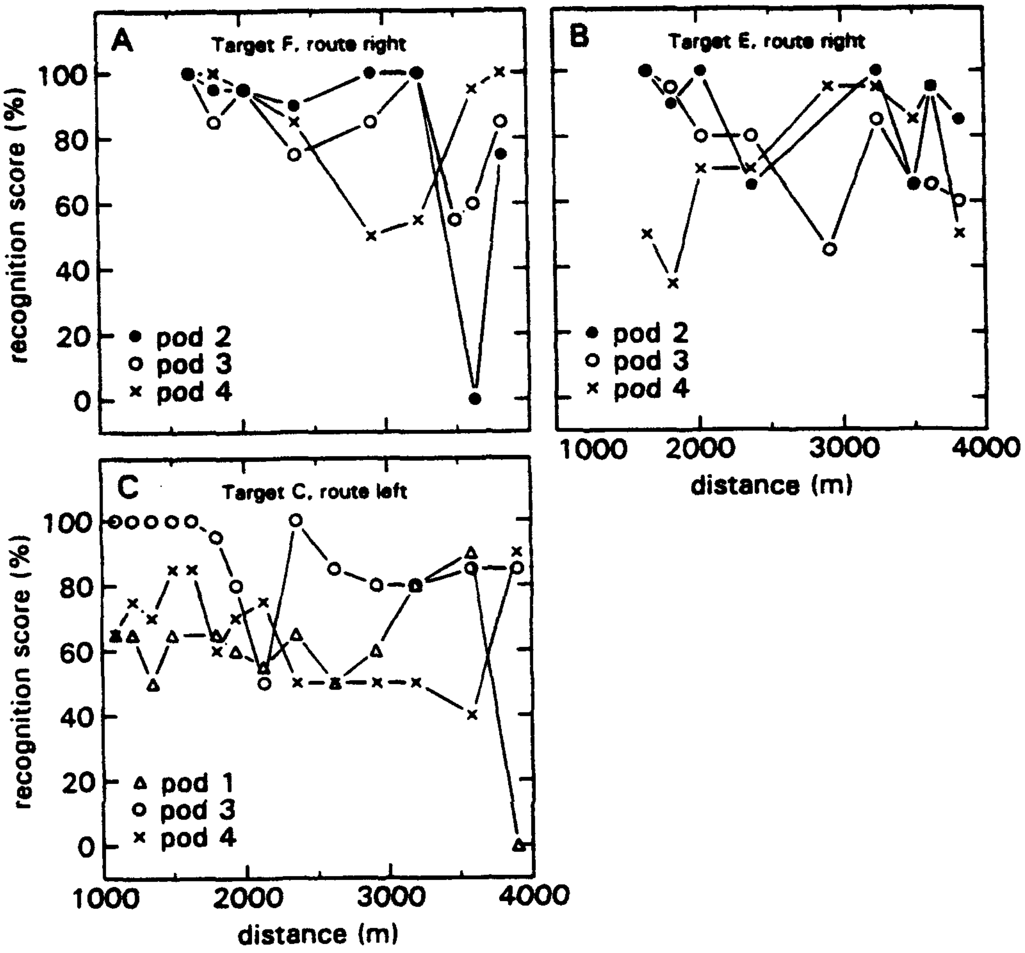


Fig. 2 Recognition performance as a function of target distance for different PODs. a: target F, route right, POD 2, 3 and 4; b: target E, route left, POD 2, 3 and 4; c: target C route left, POD 1, 3 and 4. See text for details.

A comparison between morning and afternoon (POD 2 and POD 3) is possible for five different conditions (two from Experiment 1 and three from Experiment 2). For two conditions, performance was almost identical; two conditions yielded slightly better scores for POD 2 (these conditions are shown in Figs 2a and b), and in one condition the scores for POD 3 were higher. Statistical analysis showed that there is no significant main effect ($P > 0.05$) between the scores for POD 2 and POD 3, but there is an interaction between POD and condition. This means that the overall score is similar for the two day parts, although for some runs performance may be different for morning and afternoon.

We further analyzed the differences between daytime (POD 2 and 3 were combined) and nighttime (POD 4) performance. Comparison was possible for six conditions from Experiment 2. Again, no significant main effect was found, but there was an interaction between day/night and condition. The late night (POD 1) was not included in the analysis, because only two late night runs were available in the observer experiments.

The conclusion is, that there is no part of day on which overall performance is better or worse than on other day parts. However, one vehicle may be recognized better during the morning, and another during the early night. Such an interaction is not unexpected and may be closely related to circumstantial factors, like differences in contrast between target and background.

4.3 Approach route

A comparison between acquisition performance for the left and right approach route is possible for six conditions: three for which the part of day is the same, and three for which it is different. The results of the comparison are illustrated in Figs 3a-c, where recognition scores are plotted as a function of target distance. Filled circles represent the data for a target on the left approach route; open circles for the target on the right route.

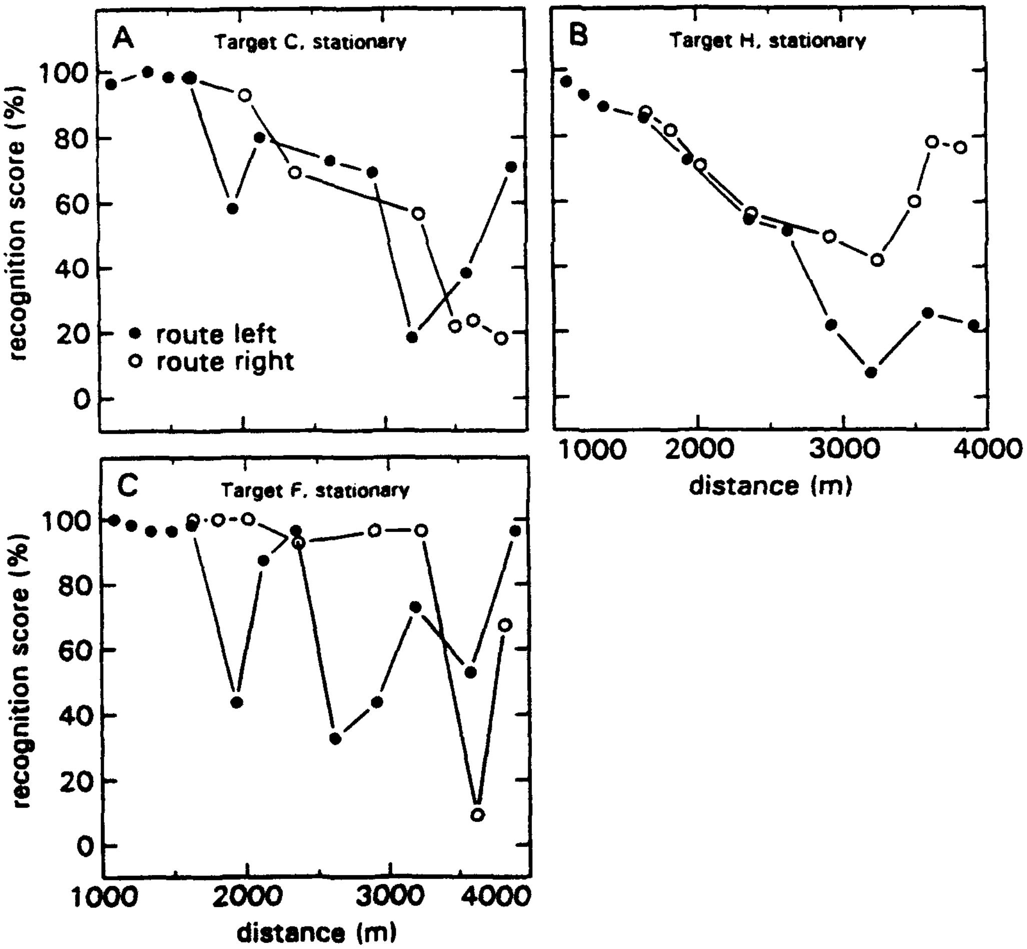


Fig. 3 Comparison between acquisition performance for the two approach routes. a: the overall course of the two curves is similar; b: performance is identical for the two routes at close range, but significantly better for the right route at large distances; c: target terrain interactions cause large performance differences for the two routes.

In Fig. 3a, the overall course of the two curves is similar, which means that there is no important difference in performance for the two routes. Fig. 3b shows that recognition performance for the target H is identical for the two routes at close range, but significantly better for the right route at large distances. Finally, in Fig. 3c, large terrain effects occur on the left route, whereas target F can be recognized easily on the right route. Thus, depending on the circumstances, performance for two routes, separated by a relatively small distance (60 - 200 m), is sometimes similar, and sometimes very different. One of the reasons might be that the left route was used more often, and was clearly visible as a white, hot track during the last days of the trials. These results again illustrate the importance of the local background structure and its interaction with target signature.

4.4 Thermal camouflage

In Figs 4a-f, recognition scores for the camouflaged targets A, D and G are compared to those for the uncamouflaged versions of these vehicles. Fig. 4a is obtained from Experiment 1; all other results are from Experiment 2. Figs 4a, b, d and f were recorded on the same day and part of day; Fig. 4c on the same part of day, and Fig. 4e on different parts of day. It should be noted that the observers were not trained on camouflaged vehicles. Figs 4a and 4b show that for target A, camouflage does not influence acquisition performance greatly; in Fig. 4a the camouflaged version is recognized slightly better than the uncamouflaged vehicle, and Fig. 4b again illustrates the effects of target terrain interaction, but shows that the overall performance is similar for both the camouflaged and uncamouflaged targets. The same findings hold for the target G (Figs 4c and 4d). There is, however, a large difference in performance for target D (Figs 4e and 4f). In most situations, this target is recognized very well for ranges up to 3500 m. Recognition of the camouflaged vehicle, however, breaks down at about 2000 m. This breakdown for target D was found on all material available (four runs). Analysis of the observer responses showed that target D was often confused with target E. Thus, with respect to target recognition, thermal camouflage was not effective for targets A and G, but very effective for target D.

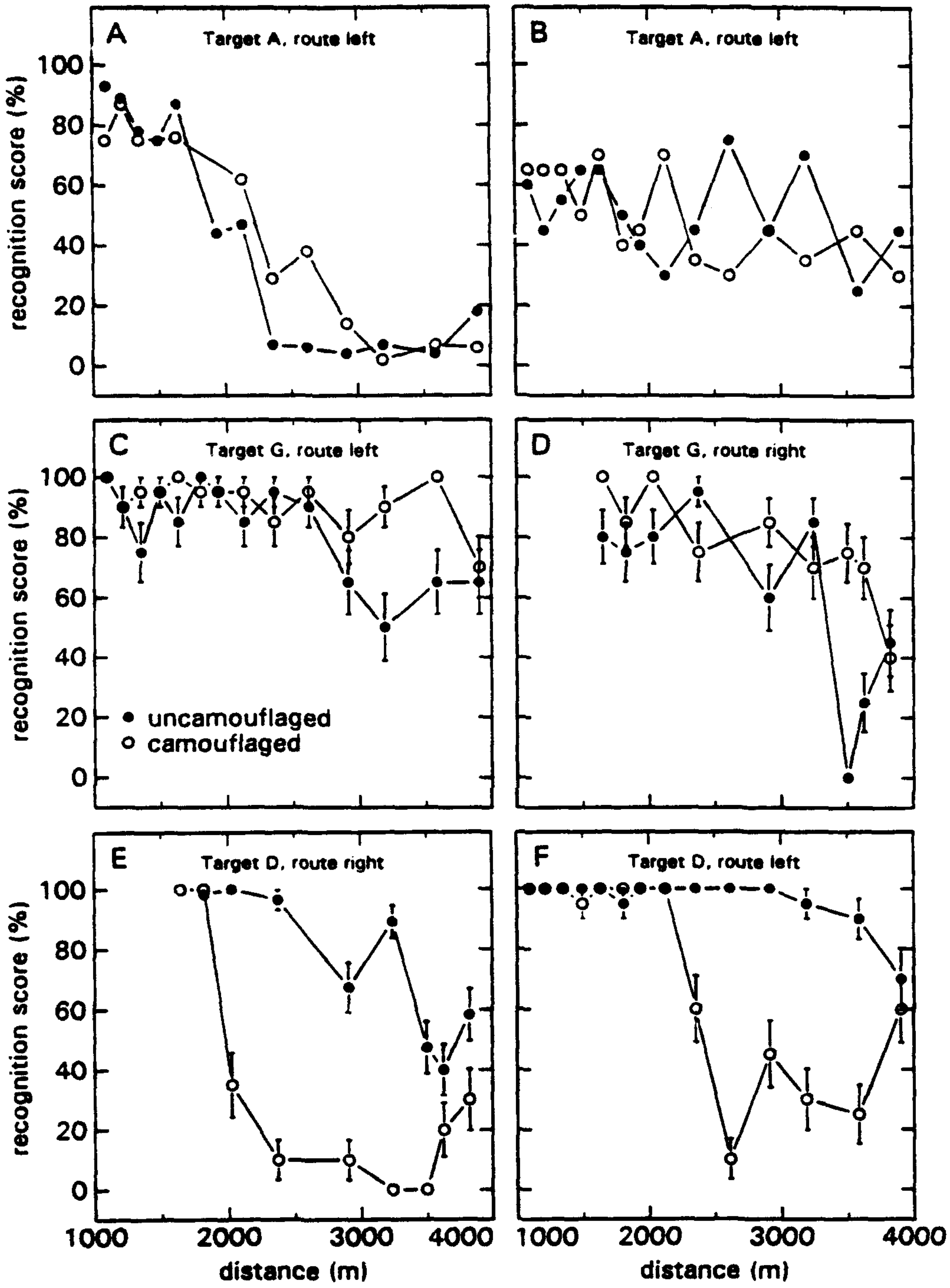


Fig. 4 The effects of thermal camouflage. Filled circles: uncamouflaged vehicles; open circles: camouflaged vehicles. No differences in performance are found for target A (a and b), or for target G (c and d). The camouflage of target D (e and f) is very effective.

4.5 Target motion

A direct comparison of performance for stationary and moving targets was possible for four conditions. The targets were approaching along the left route, and all recordings were made during the morning (POD 2) of two days.

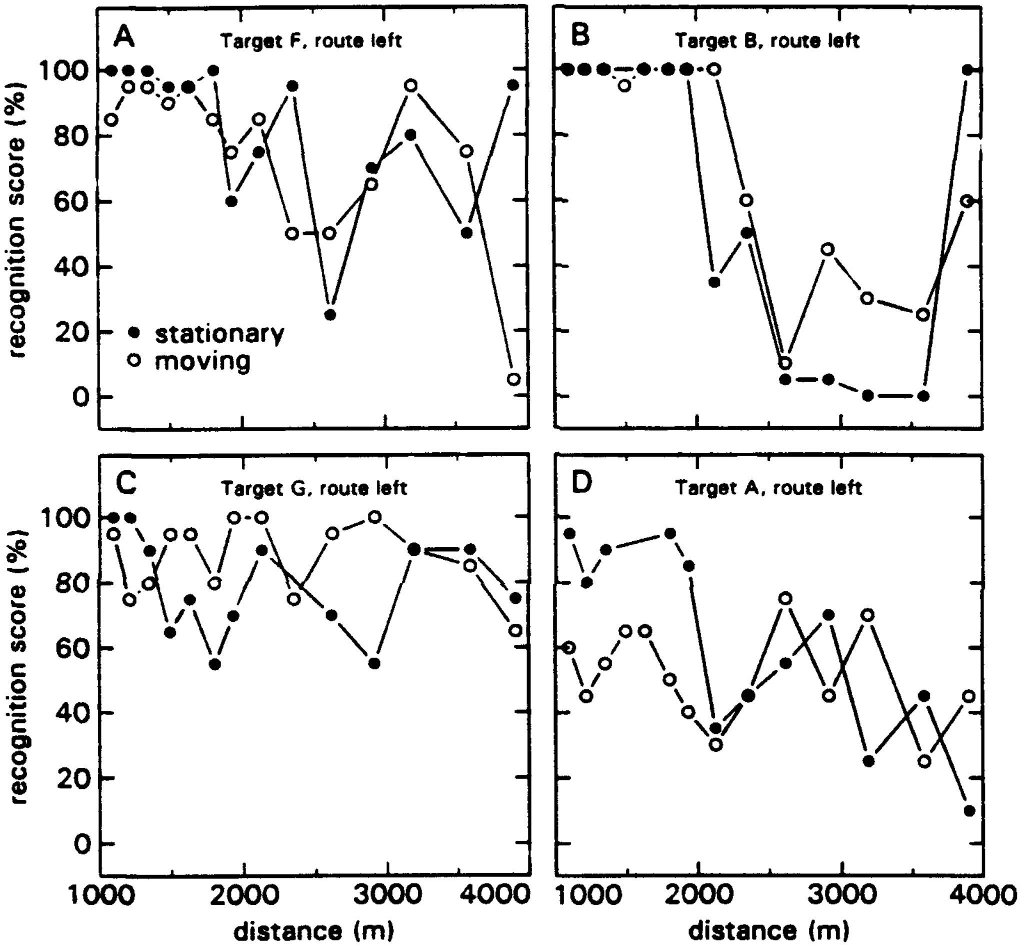


Fig. 5 Recognition scores as a function of target distance for stationary (filled circles) and moving (open circles) targets. No overall effect of target motion is found.

In Figs 5a-d the recognition scores for these targets, obtained in Experiment 2, are presented. Fig. 5a shows that for target F the differences are small, and mainly due to target terrain interactions; at some positions the moving target is recognized better, at other positions the stationary vehicle. The moving target D (Fig. 5b) is slightly better recognized than the stationary one at long range. A slightly better overall performance is obtained for the moving target G (Fig. 5c). Finally, Fig. 5d shows that there is no difference in performance for the stationary and moving target A at large distances; at short range, however, the recognition score for the stationary target is much higher.

The differences in identification and recognition performance for stationary and moving targets are small, but statistically significant ($P < 0.05$). These differences are entirely due to the lower score for the moving target A at short range (Fig. 5d): if these data are excluded from the analysis, no statistically significant differences between moving and stationary targets remain. It is uncertain that the low recognition score for the target A is due to target motion per se; other factors may influence performance as well. Inspection of the imagery showed that there were no dust clouds near the target. However, the contrast between target and background was much lower for the moving vehicle than for the stationarv target. This complicated recognition seriously.

5 DISCUSSION

Acquisition performance was determined for a large number of runs under various conditions. In the present report, we concentrated on the effects of several parameters on acquisition performance.

A major outcome of the experiments is that, in a number of conditions, strong undulations in the relation between target distance and acquisition performance are found. This may be ascribed to the interaction between a target and the local background properties ("target terrain interaction"). A strong target terrain interaction is found, e.g., if the contrast between target and background is relatively low. In that case local changes in background temperature or texture may change the contrast and apparent shape of the target considerably. Terrain interactions complicate the comparison of performance for different conditions. For example, performance for comparable runs along two routes, separated by a relatively small distance, is sometimes similar, and sometimes very different.

Current acquisition models predict a gradual, monotonous relation between target distance and performance, like the curves in Figs 1a-d. These predictions are based on parameters like resolution, target size, mean target contrast and global terrain properties. Actually, the models do not distinguish between a textured and a uniform background. In order to predict the effects of target terrain interactions (see e.g. Figs 1e and 1f), local contrast and local background properties have to be taken into account.

The influence of target terrain interaction is especially apparent in the results for pop-up targets (position presentation). In that case, the only information available is that of the target at the present position. During a target approach (sequential presentation), information from earlier target positions may be used if the present information is poor. Therefore, the sequential presentation order yields more stable results.

Performance for a target approach can be predicted from the data obtained with the position presentation. The most simple procedure to obtain an approximation of the score at a certain position, is to take the highest score for the pop-up targets at all positions between the largest and the actual distance. In such a model, performance increases monotonically as the target approaches. In a more refined model, the amount of information transfer from each presentation is taken into account to predict the choices of an observer during a target approach. This amount can be derived from the confusion matrix (see Wagenaars, 1990). For example, if the responses to an image are distributed evenly over the response categories, the observer just guesses and the information transfer is low. If, on the other hand, the observer consistently makes the same choice (this choice may be incorrect!), the information transfer of an image is high. The image that gives the largest information transfer, contributes most. The advantage of the second procedure is that dips in performance may occur at certain positions, and that it will predict smoother curves than the first procedure. Both phenomena are indeed observed in Fig. 1.

There are no indications that overall acquisition performance is different between morning (POD 2), afternoon (POD 3) and early night (POD 4). This means that the amount of clutter, which varied during the day, does not have a large influence on identification and recognition scores. For individual runs, however, performance may be different on different parts of day.

The thermal camouflage did not reduce the recognition scores for targets A and I, but it was very effective for target D. The camouflage mainly consisted of nets that covered the hot spots (like the engine) of the vehicles; the shape of the targets was not changed significantly (Clement, pers.comm.). At the front side of the targets A and G, there are no hotspots except the tracks. This explains why camouflage is ineffective for a front-view of these two vehicles.

Head-on target motion did not have a large influence on overall identification and recognition performance. A small (negative) effect was found, but this was entirely due to a single run in which the contrast between target and background was very low. A similar finding was reported by Vonhof and Rogge (1992), and by Wester and Van de Mortel (1990), who analyzed observer responses collected during the field trial. Thus, if acquisition models are extended to target motion, head-on motion may be treated as the static situation with regard to identification or recognition. With respect to search and target detection, motion will probably have a large effect.

6 CONCLUSIONS

1 Observation experiments were carried out with thermal images of stationary and moving single target vehicles. Identification and recognition performance was determined for a large number of runs under various conditions.

2 Head-on motion of targets does not have a large influence on identification and recognition performance.

3 Acquisition performance was similar for morning, afternoon and early night runs.

4 In a number of runs, performance was largely influenced by the interaction between a target and the local background properties ("target terrain interaction"). These effects are not predicted by current target acquisition models.

5 If target terrain interactions play a role, there is a large difference in acquisition performance for a target that suddenly appears, and an approaching target that has been spotted for a while. Performance for approaching targets may be predicted from the results for pop-up targets.

CHAPTER 5

STUDY II: THE RELIABILITY OF OBSERVER RESPONSES

SUMMARY

In July and August 1990 the field trials BEST TWO (Battle field Emissive Sources Test under european Theater Weather and Obscurants) were held in Mourmelon in France. Thermal images of single target vehicles were presented to observers in the laboratory to determine target acquisition performance. In these experiments observers were asked to give *two* responses after each presentation: first they were forced to name the target, even if they were not sure which vehicle was presented. Second, they were asked to indicate whether they were able to identify (I), recognize (R) or only detect (D) the target. The first (forced-choice) procedure has the advantage that performance is not biased by observer confidence, and this procedure appears to yield the maximum score that can be obtained. With the second answer, observer performance is obtained for free—or unforced—identification and recognition reports, and this procedure is more similar to the target acquisition task in a practical situation. These scores appear to be partly determined by the observer's confidence. The difference between forced and unforced responses gives a direct indication of the influence of observer *behavior* on target acquisition performance. We also determined the reliability (= probability of correctness) of first, unforced I- and R-reports during a target approach. The influence of observer behavior, acquisition level (I or R), target distance, target type, part of day and approach route on these reports was analyzed. The implications for target acquisition modelling will be discussed.

1 INTRODUCTION

Thermal imagery, collected at the BEST TWO field trials (France, July-August 1990), were used in an (observer) laboratory experiment in order to obtain identification and recognition scores for single target vehicles. In the three previous chapters we described the experimental method and presented recognition and identification performance under a wide variety of conditions. All responses were obtained with a forced-choice procedure, i.e. the observers were forced to name the target, even if identification or recognition is practically impossible (e.g. because the target vehicle is too far away). The reason to use such a procedure is that it yields objective scores (which means that performance is not biased by the subjective confidence of an observer). However, the procedure differs fundamentally from the task of a military observer in a practical situation: if an observer in the field sees a (e.g. approaching) target, he will usually first report a detection, and then, *only if he is quite sure*, report a recognition or an identification. Since the observer is free to wait until he feels sure, we will call these reports unforced. The quality and the number of unforced reports may depend strongly on - subjective - observer confidence and instructions. If, for example, an observer shows a conservative behavior, targets will be at close distance before he is confident enough to give a report. The reliability of these reports will be relatively high (there are few false alarms). On the other hand, if the observer shows a lenient behavior, his reports will be earlier (while the targets are still far away), of course at the expense of their reliability. Thus, both observer task and observer behavior may influence the outcome of an experiment.

In psychophysical research, the importance of measurement procedure and observer confidence has long been realized, and various methods have been developed to separate performance from observer bias. An overview of these methods is given by Bartleson and Grum (1984). In target acquisition research, however, the influence of these factors on performance has never had much attention. In a recent paper, Sanders et al. (1991) introduce Signal Detection Theory (one of the psychophysical methods described in Bartleson & Grum, 1984) as a tool to determine acquisition performance free from observer bias, and re-estimate some of the Johnson criteria (Johnson, 1958). Unfortunately, they did not repeat the experiment with the original procedure used by Johnson. This would have made possible a direct comparison of the results obtained with the two procedures. Moreover, it is important to keep in mind that observer bias evidently plays a role in the practical acquisition task. It is not useful to eliminate this factor by using an advanced psychophysical procedure, if we do not also determine how large the influence of observer behavior actually is. If the effects are large, this may have an important impact on, for instance, target acquisition modelling.

In order to simulate the practical situation, in the experiment the observers were not only forced to respond, but they were also asked to qualify each response as either an identification (I), a recognition (R) or a detection. The I- and R-reports can be regarded as equivalent to unforced identification and recognition reports in the field. Thus, in a single experiment, we obtained both objective and subjective scores. These scores can be compared directly. Furthermore, we are able to quantitatively determine the effects of (differences in) observer behavior on target acquisition performance.

We also determined the reliability of first I- and R-reports for approaching target vehicles under a variety of conditions. A first report is decisive, for instance, in the case that a gunner is instructed to fire as soon as he recognizes or identifies a specific target. A wrong judgement leads to waste of ammunition, and may be dangerous. It is of obvious importance to know the reliability of his initial judgement, and to obtain insight on the factors influencing reliability. Such knowledge can also be of use for a commander, who may have to make a decision based on several reports from different observers.

2 METHODS

2.1 General

The experimental method is explained in detail in a Chapter 3. Thermal images, containing one of nine single target vehicles, were presented to the observers for 5 seconds. The target vehicles are listed in Table II of Chapter 2.

Target distance varied between 4000 m and 1000 m. The targets were presented at successive positions, thus simulating a target approach (a run) along one of the two routes. Table I shows an example of a run. In the two leftmost columns, the presentation numbers and the corresponding target distances are given. For example, in the first presentation of this run, target distance is 3900 m; in the next presentation it is 3583 m. A run consisted of 10 - 15 stop positions.

Only daytime runs (recorded during the morning or the afternoon) of stationary targets were used. The total number of runs was 15. Each run was repeated 3 times. Thus, for each observer, a maximum of 45 first R- and I-reports may be obtained.

Eight male civilian and seven military observers participated in the experiment. As was shown in Chapter 2, 11 of these observers were selected on the basis of a predefined criterion.

2.2 Observer task

The observer's task is to name the target, and hit the designated key on a response panel after each presentation. In addition, the observer is asked to qualify his response as either an I (identification), R (recognition) or D (detection only). For example, if an observer is sure that the presented target was a Leopard 2 tank, he presses Leopard 2, followed by I. If he is sure that it was a tank (either a Leopard 2 or an AMX-30), he presses Leopard 2 or AMX-30, followed by R. If he thinks it's an AMX-30 (tank) or an AMX-10 (APC), he presses AMX-30 or AMX-10; D, because he is not able to distinguish between two different vehicle classes.

I-responses can be regarded as equivalent to (unforced) identification reports in the field. R-responses are equivalent to recognition reports (e.g. Leopard 2; R is equivalent to reporting a tank in the field). D-responses are not analyzed, because the experiment was designed in such a way that detection of the target was always possible.

2.3 Analysis

For each run the responses were analyzed in the following way. Table I shows a simulated run of a Leopard 2 tank. The two leftmost columns give the presentation number and the target distance. The middle column of Table I shows the subject's responses to the successive presentations of the approaching target. The two rightmost columns give the analysis of these responses in terms of a forced and unforced report, respectively.
For example, at the first presentation the Leopard is at 3900 m distance. The response of the observer, PRI, is of the wrong vehicle type and class. Thus, the forced response is neither a correct identification nor a correct recognition. The qualification "D" means that, at this stage, the observer is not confident enough to report an identification or recognition.

Table I Simulated of a run of a Leopard 2 tank, approaching along the left route. See text for details.

pres. #	distance (m)	response		forced ident./rec.	unforced report
1	3900	PRI	D	wrong/wrong	none
2	3583	Truck	D	wrong/wrong	none
3	3188	Truck	D	wrong/wrong	none
4	2913	PRAT	D	wrong/wrong	none
5	2615	Leopard 2	D	correct/correct	none
6	2353	AMX-30	D	wrong/correct	none
7	2124	AMX-30	R	wrong/correct	correct rec.
8	1933	AMX-30	R	wrong/correct	correct rec.
9	1631	AMX-30	I	wrong/correct	incorr. ident.
10	1494	Leopard 2	I	correct/correct	correct ident.
11	1349	Leopard 2	I	correct/correct	correct ident.
12	1215	Leopard 2	I	correct/correct	correct ident.
13	1096	Leopard 2	I	correct/correct	correct ident.

If all responses, regardless of the acquisition qualification, are taken into account, forced-choice scores are obtained. In Table I, identification is correct for distances below 1494 m and, probably due to a lucky guess, at 2615 m. In a similar way, recognition

scores can be obtained: under forced conditions the Leopard 2 has correctly been recognized as a tank for all distances below 2615 m.
In order to obtain identification and recognition probabilities, for each image the number of correct identification and recognition responses are divided by the total number of presentations (averaged over observers and repetitions). Forced-choice performance is discussed in Chapters 3 and 4.

At distances greater than 2124 m, no R- or I-report has been made. This means that the observer is not sure enough to pass on information about the vehicle class or type. Thus, in the example of Table I, (unforced) correct recognitions may only occur for distances ≤ 2124 m. From this distance on down, all R-reports happen to be correct (i.e. the Leopard 2 has been correctly recognized as a tank). Similarly, an (unforced) correct identification may only occur for distances ≤ 1631 m. The first I-report is incorrect; at nearer distances, all I-reports are correct. Identification and recognition probabilities are obtained by dividing the number of correct I- and R-reports by the total number of presentations (averaged over observers and repetitions).

The first R-report occurs at a distance of 2124 m, and happens to be correct in this example. The first I-report, at 1631 m distance, is incorrect.
The reliability, or percentage correct, of first reports is defined as the number of correct first I/R-reports, divided by the total number of first I/R reports.
An R- or I-report, given at the largest distance (the first position of a target), strictly cannot be treated as a first report, because recognition or identification might be possible at (much) longer distances.

3 RESULTS: FIRST REPORTS

3.1 Observer differences

For each observer we determined the overall percentage correct of first recognitions and identifications (averaged over all runs and repetitions). The results are presented in Table II.
At both acquisition levels, the mean overall score is about the same: 75% vs 80%.
The overall variation (± 1 s.d.) in reliability is 8% for identification and 12% for recognition. Individual differences may either be due to differences in observer conservatism (some observers wait a little longer before they report an R or I than others do) or in overall skill of the observers. As was shown in Chapter 3, there are large individual differences in the scores at short ranges, which were taken as a measure of observer skill.

Table II Overall reliability (percentage correct) of first recognitions and identifications of civilian and military observers.

observers	recognition (%)	identification (%)
civilian		
JB	81	85
EM	82	88
LO	79	93
WH	97	72
PBR	67	77
CK	84	79
mean	82 ± 9	82 ± 8
military		
JS	64	71
RC	87	70
WK	70	79
MW	67	83
PH	55	75
mean	69 ± 12	76 ± 6
overall mean	75 ± 12	80 ± 8

Possibly, the best observers also score highest on overall reliability. To test this hypothesis, we calculated the correlation between the percentage correct of first identifications (from Table II) and the percentage of correct (forced) identifications at short distances. This correlation is low (r = 0.28) and is not significant (n = 11, P > 0.05). Similarly, no significant correlation (r = 0.16) was found between the recognition scores at short distances and the score on first recognitions. We conclude that the variation in reliability is due to differences in conservatism of the observers (see also section 3.2). The overall score for civilians is higher than for military observers (13% higher for recognition and 6% higher for identification). Since both groups score equally well under forced conditions (see Chapter 3 and 4), this means that civilians act more conservative than military observers.

3.2 The effect of target distance and route

Fig. 1 shows the number of first recognitions and identifications as a function of target distance, expressed as the percentage of the total number of runs; the percentage correct of these first recognitions and identifications are presented in Fig. 2. In the figures the results for the left and right route are plotted together. Largest distance for the left route is 3900 m; for the right route it is 3820 m. As was argued in section 2.3, scores for these distances should be treated separately.

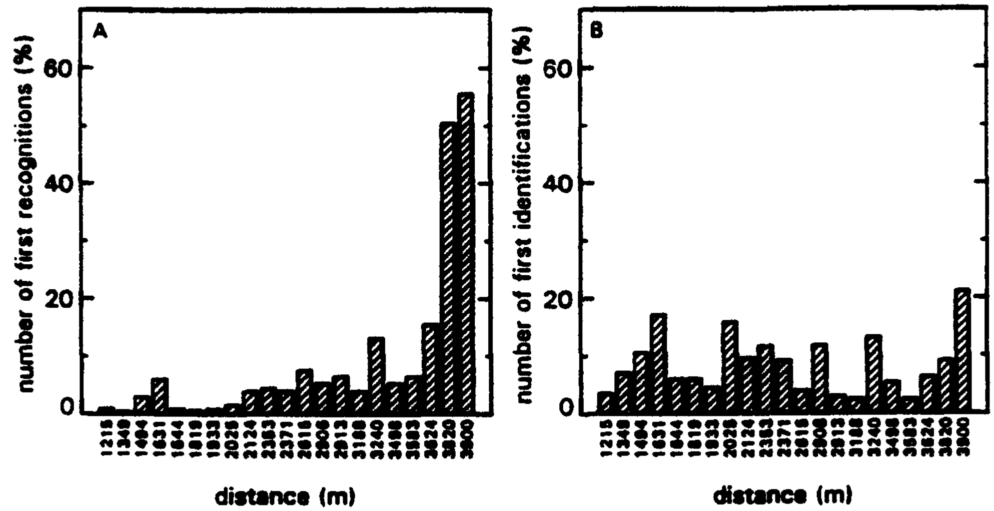


Fig. 1 Number of first reports as a function of target distance, expressed as the percentage of the total number of runs. (a) Recognition. (b) Identification. Most of the first recognitions are reported at the largest distances; the number of first identifications is distributed more evenly over the entire range.

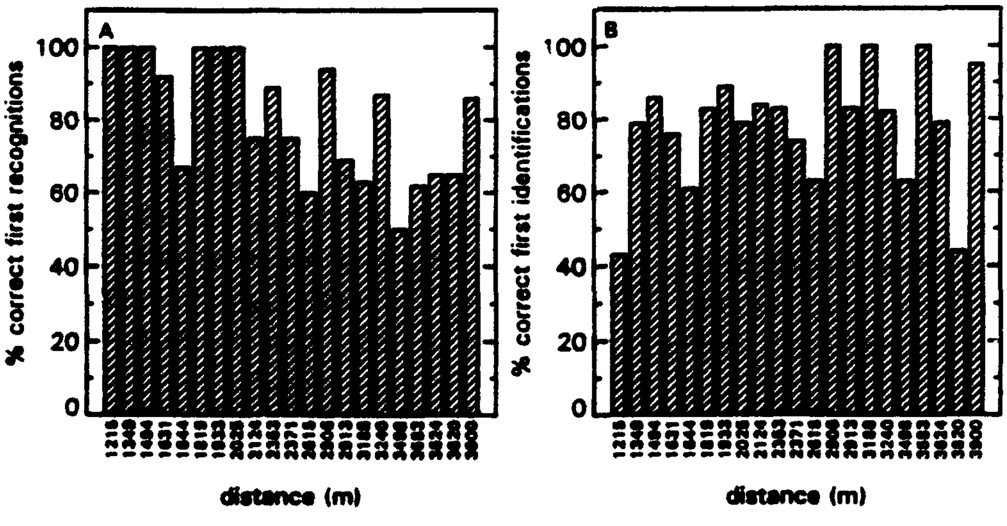


Fig. 2 Percentage correct of first reports as a function of target distance. (a) Recognition. (b) Identification. The percentage correct of first recognition reports decreases with distance. For identification, the percentage is quite independent of distance. See text for details.

Fig. 1a shows that most of the first recognitions (53%) are reported at the two largest distances. About 75% of these recognitions are correct (Fig. 2a). This means that, for our camera and under the prevailing conditions, recognition of targets at a distance of about

4 km or more is quite possible (more than half of the recognitions can be made at distances larger than about 4 km!). On the other hand, the number of first identifications (Fig. 1b) made at the largest distances is not higher than at short distances. If we omit the largest distances, we see that most of the first recognitions are reported between 2200 m and 3600 m; the number of first identifications is distributed evenly over the entire range. Only a few times, no R-report was given at all. In about 10% of the runs, no I-report was given.

Fig. 2a shows that the percentage correct of first recognitions gradually decreases with target distance. On the other hand, the mean identification score (Fig. 2b) slightly *increases* with distance. Calculation of weighted regression lines yields the following results:

- between 1000 m and 4000 m, the increase of the identification score with distance is about 10%. The lines are similar for left and right route.
- over the entire range, recognition scores for the right route are about 10% higher than for the left route. Mean recognition score decreases from about 90% at short distances to 60% at the largest distances.

The slight increase of the identification score with distance is considered not relevant. The significance of the decrease in the mean recognition score with distance is tested below.

3.2.1 *Interaction between observer behavior and distance of first reports*

The decrease of the recognition score with distance might be caused by a possible interaction between observers and the distance at which the R-reports are made: especially for recognition, large intersubject differences were found in mean overall score (see Table II). If two observers are at the same level of skill, the more conservative observer will give his first R-report at shorter distances, and yield a higher score. Consequently, higher scores are expected at shorter distances due to differences in conservatism of the observers.

In order to test whether the decrease in score with distance is real, or due to differences in observer conservatism, we divided the observers into two groups: one for which the recognition scores in Table II are better than 75% (6 observers) and one for which they are below 75% (5 observers). Second, the distances were divided into two ranges: above and below 2500 m. Then it can be tested whether the overall distance effect is present for both groups or is the result of unequal proportions of both groups in the two distance classes. The test showed that 85% of the variation may be ascribed to the observer groups. The remaining 15%, due to target distance, is not statistically significant. Thus, the decrease of the percentage correct of first recognitions with distance, shown in Fig. 2a, is almost entirely due to differential effects of observer conservatism. The test was further

extended to determine the influence of approach route (left or right). Differences due to approach route turned out to be not statistically significant.
We conclude that target distance and approach route do not significantly influence the reliability of first R- and I-reports.

3.3 The effect of Part Of Day

The experiment was carried out for two parts of day: morning and afternoon. Mean identification scores are 78% and 81% respectively, and recognition scores are 78% and 69% respectively. An extension of the test, introduced in section 3.2, showed that the variation in recognition scores due to part of day is significant ($P < 0.05$), but small compared to the variation caused by observer differences.
We conclude that there are no relevant differences in reliability due to part of day.

3.4 The effect of target type and camouflage

No significant differences in the reliability of first R- and I-reports were found between target types. The results, however, show a slightly lower score (10-15%) for two camouflaged vehicles than for the uncamouflaged versions. This difference is considered not relevant because it is smaller than the intersubject differences.

4 RESULTS: FORCED VERSUS UNFORCED RESPONSES

In this section, the probability of an observer giving a correct identification response under forced-choice will be compared to that of giving a correct unforced I-report. In Fig. 3a-d, identification scores are plotted as a function of target distance for four different runs. Open circles represent unforced responses, filled circles represent forced responses.Roughly, three different cases can be distinguished:

1 If the task is relatively simple (i.e. at near distances), differences between the curves are small. In this case, observers are very confident, and most of the responses will be I-reports. Therefore, no differences are expected.
2 In a more difficult situation, the observers become less sure of the vehicle type, which means that less (unforced) identifications are reported. If, however, the observer is forced to make a choice, identification performance is good, as is shown by the upper curves in Fig. 3a-c (at distances between 2000 m and 4000 m). In this case, there are large differences between the scores for the forced and the unforced task.

3 If the task is very difficult (Fig. 3d), there will be no I-reports at all. With the forced-choice procedure, probability will be at guess level (which is 17% for six different response categories).

Thus, for most runs, at intermediate distances (case 2) the identification score obtained with the forced-choice task is much higher than it is if the observer is free to respond. This means that, although the observer is not confident enough to give an I-report, his responses are still quite reliable. As a consequence, identification *ranges* (which may be defined as, for instance, the ranges where identification performance is better than 70%), may differ by more than a factor 2 in the examples of Fig. 3.

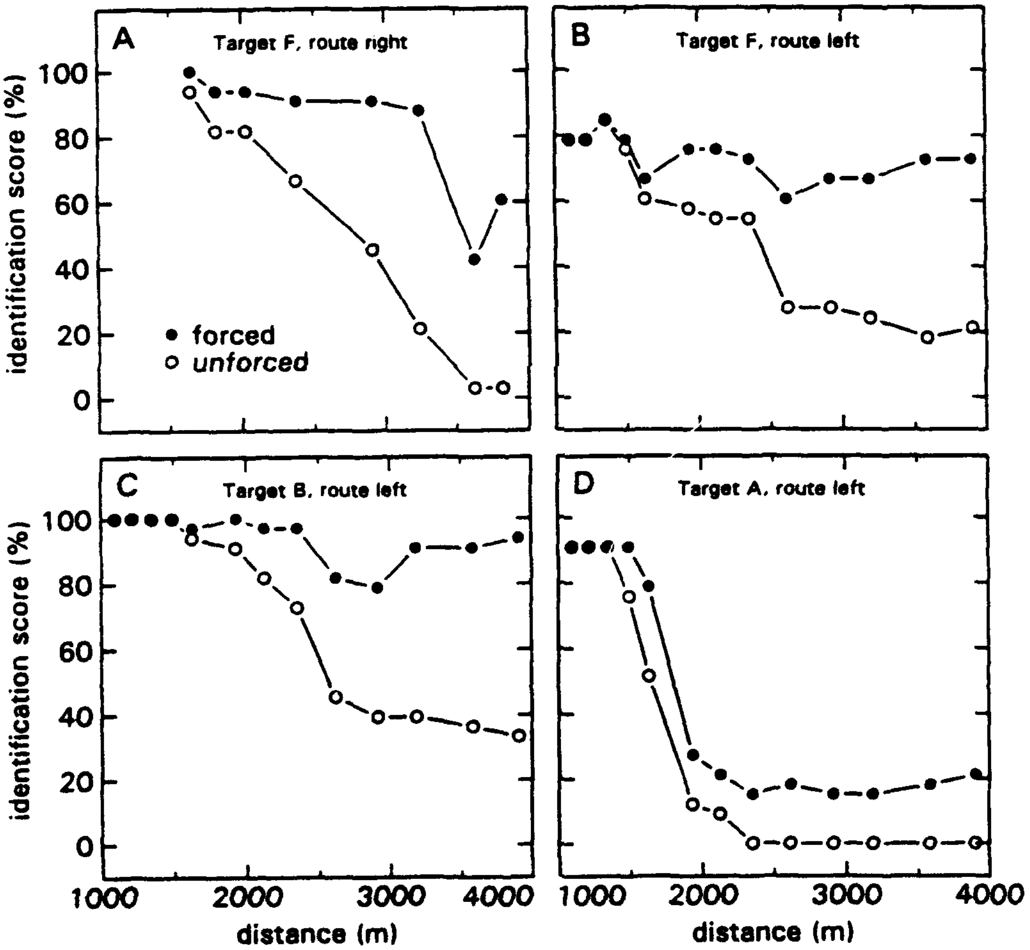


Fig. 3 Identification scores as a function of target distance for four different runs. Open circles: unforced responses; filled circles: forced responses. See text for details.

5 DISCUSSION AND CONCLUSIONS

In the present target acquisition experiments, identification and recognition scores were obtained with two different measurement procedures: one in which the observer was forced to give a response, and one in which he was free to wait until he felt sure. The results of these tasks can be compared directly.
A forced-choice task has the advantage that it yields objective scores, that are not biased by a subjective confidence criterion of an observer. The unforced task, however, is more similar to the actual task of an observer in a practical field situation. Moreover, the chance level, which can be considerable in a forced-choice task with a limited set of targets, has been reduced or eliminated. The score, however, depends on a subjective criterion of the observer, which in turn can be affected by the instructions given to him (see below).

The probability that a first, unforced R- or I-report is correct, is about 75%-80%. This percentage is rather independent of factors that influence acquisition performance, such as target distance, target type, camouflage, part of day or approach route. A similar finding was reported by Vonhof and Rogge (1992), who analyzed observer responses, collected during the BEST TWO field trial. They reported that the reliability of first I- and R-reports is about 80%, regardless of the target distance (between 1000 and 4000 m). At the same time, mean acquisition performance varies greatly over that range. In their experiment, both task and, e.g., camera type used were different from those used in our experiment. The results suggest that the reliability of first reports is determined by a fixed internal "risk criterion" of the observer, which is independent of the outside conditions. An observer will only decide to give a report at a higher acquisition level if he thinks that the risk is acceptable (i.e. below his internal criterion). If this hypothesis is correct, the finding will not only be valid for the restricted conditions under which the BEST TWO trials were carried out, but may be considered as a more general rule.

The present results, however, also show that different observers show a considerable difference in risk criterion, especially for the recognition task: mean correct scores vary between 55% and 97%. This criterion is not correlated with observer skill (i.e. the ability of an observer to correctly identify or recognize a target if he is forced to). Thus, some observers are more conservative than others, wait longer before they give a report and, consequently, yield higher scores. Surprisingly, civilian observers turned out to be more conservative than military observers.

The two measurement procedures we used, yield important differences in the probability of a correct response. Identification ranges, obtained with the forced-choice procedure, are usually much longer than those obtained if the observers are free to give a report. This difference may be ascribed to observer conservatism. The largest differences in score arise in those situations where the observer is not sure, but chooses correctly if he is forced to. The fact that those differences are considerable means that the observer possesses more information than he actually uses. The risk of making a mistake prevents him to give a report.

A practical consequence is that the instructions to an observer may have a large impact on the information he will pass on. Therefore it is advisable that a commander adapts his instructions to the actual situation in the battlefield. If wrong reports are very dangerous, he will instruct the observers to be very conservative, and he will obtain relatively few reports, which will be of high quality. If he wishes to have more information, he will ask the observers to try to identify the targets as soon as possible. This may extend the acquisition ranges considerably, of course at the cost of a higher false alarm rate. The forced-choice procedure corresponds to the actual limit of unconservative behavior.
The present finding also has important implications for target acquisition modelling. The forced-choice score is the maximum score that can be obtained, given the contrast, the resolution, the capacities of the human visual system and the chance that an observer guesses correctly. Therefore, the forced condition may be regarded as an important condition for target acquisition modelling. Observer conservatism acts as a filter that describes the difference between the forced condition and a practical situation.

We suggest that a target acquisition model should consist of two stages: the first stage calculates the maximum score that may be obtained under the circumstances. This score is only limited by the physics and physiology of the systems involved, and is independent of observer behavior. In fact, current acquisition models are designed according to this principle. In the next stage, the information is passed through the "uncertainty" filter, and the actual probability of an unforced report is calculated. The filter characteristics depend largely on observer confidence. From the present experiment, filter characteristics, and their variation due to differences in conservatism, can be determined. The characteristics will also depend on the instructions to the observer. More research is required if we want to determine the impact of this factor.

CHAPTER 6

EVALUATION OF TARGET ACQUISITION MODEL "TARGAC" USING "BEST TWO" OBSERVER PERFORMANCE DATA

SUMMARY

The TARGAC model predicts target acquisition performance for a variety of sensors in the visible and thermal infrared. A comparison was made between TARGAC recognition performance predictions and measured observer performance for a large number of trials, using thermal imagery collected during the BEST TWO field test (France, 1990). The evaluation shows that there are important differences between measured and predicted recognition performance. On average, observer performance is considerably better than the model predicts: a correction factor of 1.80 should be applied to match the recognition range predictions to the results of the experiments. Further, the model does not give accurate predictions for individual targets on specific backgrounds: the ratio between observed and predicted recognition range varies between 0.9 to 3.6 (95% criterion). The routine that describes electro-optical and human visual system performance is responsible for the predictions. This routine is based upon the widely-used NVESD Static Performance Model which uses the so-called 'Johnson Criteria'. Hence, the findings of this evaluation may also hold for other models based on these criteria. In addition, it was found that the the version of TARGAC that was tested contained a number of problems and software errors. Corrections are suggested.

1 INTRODUCTION

Target acquisition models predict how well human observers, using an optical or electro-optical (E/O) viewing device, are able to detect, recognize or identify a military target. The input variables are the properties of the target and its background, the atmospheric conditions, and the properties of the viewing device used. The output is a relationship between the distance from the target to the sensor, and a probability of correct detection, recognition or identification. Target acquisition models are used, for example, as Tactical Decision Aids (TDA's), in war games, and as a tool to compare performance of competing sensor systems for a specific task. A comprehensive target acquisition model is TARGAC, developed at the U.S. Army Research Laboratory, Battlefield Environment Directorate (ARL-BED). The model is part of the Electro Optical Systems Atmospheric Effects Library (EOSAEL).

The evaluation of target acquisition (TA) models is of interest because the reliability and accuracy of their predictions is not always known. The models are usually based on theoretical knowledge of E/O-device physics, atmospheric optics and human vision. However, target acquisition is an extremely complex process, and many processes that play an important role in visual performance, are not yet understood. Therefore, cognitive factors are often not incorporated in models, and predictions are made for artificial targets in a laboratory environment rather than real targets in the field. No one knows, however, how significant the effects of these omissions are. An important side effect is the so-called *"false precision problem"*: because the accuracy of target acquisition model predictions is unknown, they are often treated as being exactly correct. Therefore, it is necessary to measure observer performance for realistic field conditions and to evaluate model predictions empirically.

Ideally, a TA model evaluation provides a quantitative measure of the accuracy of the model predictions, and indicates for which applications the model may be used and what the restrictions are. The evaluation may also give indications for model improvement. However, the complexity of the acquisition process makes model evaluation very difficult. It is further difficult to obtain accurate and reliable observer performance measures for realistic field conditions. In the field, conditions are hard to control, and there is little or no opportunity for repeated trials under identical conditions, which are needed to obtain statistically meaningful results. Often, only qualitative conclusions can be drawn from the results of a field trial, and the results of the evaluation do not provide insight in the reliability of the model, nor indicate how the model may be improved. As a result, adequate evaluations of TA models using field data are sparse.

One of the test objectives of the NATO AC243/Panel4/RSG.15 field trial BEST TWO, held in France (1990), was to collect observer performance data with sufficient accuracy for a quantitative evaluation of TA models. During the test, a large amount of images of stationary and moving target vehicles at many distances were recorded. Thermal (8-12 μm) images were used in a laboratory experiment to measure target recognition and identification performance for a large number of observers , see previous the sections. A limited observer experiment was carried out in the field for validation of the observer scores that were measured in the laboratory.

The results of the BEST TWO observer performance experiments are used for the evaluation of TARGAC. A comparison will be made between TARGAC recognition performance predictions and measured observer performance for a large number of trials. The result will be expressed as a probability distribution of the ratio between measured and predicted acquisition ranges. The *mean* of this distribution quantitatively shows how well the model predicts *overall* acquisition performance over a large number of trials. The *variance* of the distribution is a quantitative measure of the accuracy of the model predictions for *individual* trials.

The TARGAC evaluation is carried out in five steps:

1. TARGAC predictions are calculated for the BEST TWO situation. To be able to do this, extensive meteorological data for the BEST TWO situation, data on the BEST TWO target set, and the Minimum Resolvable Temperature Difference (MRTD) curve for the thermal imaging system that was used, was collected and fed into the model.
2. A sensitivity analysis is performed because not all of the input information is available with a high degree of accuracy. This analysis shows the extent to which changes in each input parameter influence the model output. Parameters for which the model is not very sensitive need not be specified with great accuracy, while parameters for which a high sensitivity is found must be provided with high precision.
3. The TARGAC predictions are plotted as probability of a correct recognition response *vs.* target range together with the observer data. Graphical comparison gives a first impression of the quality of the predictions.
4. For each trial, the ratio between actual and predicted recognition range is calculated. A set of BEST TWO trials yields a probability distribution of this ratio. Mean and variance of the probability distribution are a measure of the reliability of the model predictions for the set of BEST TWO trials.
5. Further analysis is carried out in order to find the possible sources of the differences between observer scores and model predictions. Suggestions for model improvement, based on the results, will be discussed.

This report is organized as follows. In section 2, a short description of TARGAC is given. Section 3 contains an outline of the BEST TWO field trials and the observer performance experiments. The sensitivity analysis is presented in section 4, and section 5 gives a simple equation that describes the TARGAC predictions for the entire set of BEST TWO runs. The comparison between the TARGAC predictions and the observer performance data is made in section 6. In section 7, the variance in the probability distribution is analyzed to find its possible sources. A discussion of the results is presented in section 8, and conclusions and recommendations are given in section 9.

2 TARGAC

2.1 General

TARGAC predicts the probability of detection- and recognition of (military) targets as a function of range, for a variety of sensors. The model is freely available (as part of the EOSAEL Library) in a PC and a mainframe version. The program runs in both interactive and batch mode. An extensive overview of the model is given in the TARGAC User Guide (Gillespie, 1991). The model basically consists of three parts: an inherent target contrast module, an atmospheric effects module, and a system performance routine.

2.1.1 *Inherent contrast calculation*

The first stage of the model calculates the inherent contrast (i.e. the contrast at the location of the target) between target and background, given the characteristics of target and background, and the meteorological conditions. For visual devices, mean or area contrast is defined as the difference between mean target and background luminance, divided by mean background luminance. For the case of thermal imaging, target and background temperature are calculated by a Thermal Contrast Model (TCM2), and the inherent contrast is expressed in terms of a temperature difference. It is possible to by-pass this module and directly input the inherent contrast.

2.1.2 *Atmospheric effects calculations*

TARGAC contains an extensive atmospheric effects module. This module calculates the contrast transmittance through the atmosphere for various wave bands, based on meteorological input data. It yields the apparent contrast of a target as seen by a sensor, as a function of range.

2.1.3 *System performance calculation*

The actual probability of acquisition is calculated using the NVESD Static Performance Model (Johnson, 1958, Ratches et al., 1981), in which target acquisition performance is described using the well known 'Johnson criteria'. These criteria link target acquisition performance with the ability to resolve dark bars of a certain spatial frequency and contrast against a uniform background. For example, the model predicts a *recognition* probability of 50% if a target is at such a range that a human observer with the viewing system is just able to resolve *four* line pairs over the effective, i.e. minimum, dimension of the target. The higher the resolution of the viewing device, or, the larger the target, the longer the range at which four line pairs can be resolved. For a 50% *detection* probability, a resolution of 1 line pair across the effective dimension of the target is required, for *identification* this is 8 line pairs. Criteria exist for different levels of probability. The relationships between the number of resolvable line pairs and probability for several

acquisition levels are called Target Transfer Probability Functions (TTPF's) (Ratches, 1976). These functions have been established experimentally by averaging over many targets, target orientations and aspects. Ratches also indicates the accuracy of the criteria: the ratio between an optimistic and a conservative criterion is 3:4. For a 50% recognition probability, the four line pair criterion, used in TARGAC for thermal viewing systems, is conservative, whereas a three line pair criterion would be optimistic.
In practice, the Johnson criteria are applied using a threshold performance curve of the viewing device that gives the contrast required to resolve a 4-bar pattern as a function of spatial frequency. For visible light devices this is called the Minimum Resolvable Contrast (MRC)-curve; infrared devices are characterized by a Minimum Resolvable Temperature Difference (MRTD)-curve.

2.2 TARGAC version

The TARGAC model is 'currently defined as being in the developmental stage of software' (Gillespie, 1991). This means that regularly new versions are released. For the present evaluation, we used the PC version of TARGAC that was released in June, 1992. During the sensitivity analysis (see section 4), a number of software errors were found and most of the bugs were fixed in consultation with Dr. P. Gillespie (ARL-BED). This means however that results presented in this paper were obtained with an improved version, and not with the standard distribution version. Since the latter still contains a number of errors, we recommend to contact ARL-BED before using the program. A complete overview of the traced errors, modifications and recommendations is given in Appendix 4.

2.3 TARGAC input

For the present study, all calculations were carried out in batch mode. In this mode, TARGAC requires an input file that contains information about target, background, meteorological conditions, geographic situation, date, time, and the viewing device that is used. In the standard version, three levels of probability (between a maximum level of 90% and a minimum level of 10%) may be specified for which TARGAC predicts a detection and recognition range. This number of levels was extended to five in the improved version. In the present study, ranges were always calculated for five probability levels: 90, 70, 50, 30 and 10%.

TARGAC has 24 built-in targets and 29 background choices. Examples of targets are: a T62 and T72 tank, a ZIL truck and a BRDM-2 anti-tank vehicle. Many of the targets are available in 'off', 'idle' and 'exercised' conditions. For background, there is, for example: 'tall grass -- growing', 'dirt road', and 'coniferous trees -- dormant'. It is not possible to enter user-specified targets and backgrounds. On the basis of target and background characteristics and meteorological input, the TCM2 module in TARGAC calculates target and background temperature. There is a possibility to by-pass the TCM2 calculations and directly input target and background temperature.

In the viewing device menu, there are 14 visible sights, 4 Image Intensifiers and 5 thermal sights, and the corresponding MRC or MRTD curves are built into the program. When predictions must be made for a viewing device that is not built in, the user must specify an MRC or MRTD curve, in the form of the coefficients of a sixth order polynomial fit to the MRC or MRTD data.

2.4 TARGAC output

The program calculates detection and recognition ranges for the probability levels that are specified in the input file. Ranges are specified with a precision of 0.1 km. TARGAC also provides several results of intermediate calculation stages, such as the inherent target contrast, when the Thermal Contrast Model (TCM2) is used.

3 BEST TWO FIELD TEST AND LABORATORY EXPERIMENTS

The BEST TWO field trials, organized by NATO AC/243, Panel 4, RSG.15, were held in July and August 1990 in Mourmelon, France. The purpose of the test was to quantify the performance of electro-optical devices under battle field conditions. A comprehensive overview of the trials is reported in Chapter 2.

3.1 Observer performance data

During the field test, recordings were made of single stationary and moving target vehicles approaching from a distance of 4000 m to a distance of 1000 m. Targets were always in front view. Image sequences, recorded from a thermal (8-12 μm) thermal imager on a U-matic videorecorder, were used in laboratory experiments to measure observer performance for target recognition and identification. The experiments are described extensively in the previous chapters. For one observer, acquisition performance with the thermal imager was measured directly in the field, and the experiment was repeated with the video tapes in the laboratory for a number of observers. No significant differences were found between field and laboratory performance (Wester & van de Mortel, 1990).

In two experiments, observer performance was measured for a total of 38 different target approaches (runs). These runs differ in target type, approach route, date and time (recordings were made during day and night). Six different targets were used, three of which were camouflaged during some of the runs. Sequences of images containing a single target were presented to observers. The observer's task was similar to the target acquisition task in a practical military situation: after each presentation, they were first asked to indicate whether they were able to identify, recognize or only detect the target, after which they had to name the target. Two ways of presenting the target images were used: "pop-up" and "approaching". In the "pop-up" presentation, a randomly chosen target was presented at a random distance. In the "approaching" condition, the images from a single run were presented as an ordered sequence, simulating a target approach from 4 km on down. Search was explicitly avoided. In Experiment 1, the number of observers was 11. Each target image was presented 5 times. Performance was measured for 15 runs for both the "pop-up" and "approaching" presentation order. Recognition scores (averaged over observers and repetitions) for these runs are presented in Appendix 2 in Fig. A9 and A12 (names have been removed because some of the information is confidential). In Experiment 2, each target image was presented 5 times to 4 observers. Performance was measured for 33 runs for "pop-up" presentation (10 of these runs were also used in Experiment 1). Recognition scores for these runs are given in Appendix 2, F[illegible] and A11.

TARGAC will be evaluated against three data sets. Data set A contains observer performance data for all (38) runs, for "pop-up" targets. Data set B contains the data for 15 runs, presented as an ordered sequence. Data set C, which is a subset of data set A, contains the data for the same 15 runs, now presented as "pop-up" targets. This set will be used for a direct comparison of the results of the evaluation for the two types of presentation order.

3.2 TARGAC input data

During the BEST TWO field test, the input data for target acquisition models were collected by several participating nations. Meteorological data were gathered by the delegations of the United States, France, Germany and the Netherlands. The data are stored in the AAODL database which is maintained by ARL-BED. Dr. P. Gillespie collected the appropriate meteorological and geographical data, date and time, and composed TARGAC input files (see 2.3) for a large number of BEST TWO runs (excluding the viewing device parameters and target and background type, see below).

MRTD-measurements (horizontal and vertical) for the thermal imager including the videorecorder, were carried out by FEL-TNO (de Jong et al., 1991).

Target and background type have to be selected from the TARGAC menu. Since the targets, that were used in the BEST TWO experiments, are not part of the TARGAC menu, comparable built-in targets had to be chosen (see also 4.1). Direct measurements of target and background temperature in the field, made by the Danish delegation (Andersen, 1991), may also be used. In section 5, it will be shown that calculations for the BEST TWO targets can be made using a simplified equation that describes the TARGAC probability *vs*. range predictions for the entire set of BEST TWO runs.

4 TARGAC SENSITIVITY ANALYSIS

Not all the input data that TARGAC requires to make predictions for the BEST TWO situation, is available, or available with high accuracy. Inaccuracies in the input are of course reflected in the output of the model, and thus affect the comparison between the model predictions and the observer performance data. The model may be very sensitive to some parameters, and insensitive to others. A sensitivity analysis shows the extent to which the model outcome is influenced by changes in the input parameters. Basically, this is done by systematically changing one parameter, while keeping all the others constant. The effect of possible errors in the input data on the outcome of the evaluation can then be assessed.

4.1 Inaccuracies in the input data

Inaccuracies in the TARGAC input data for the BEST TWO situation that may influence the model predictions, are:

Meteorological data
There are (small) differences in meteorological data collected by Germany and the U.S.. Furthermore, meteorological data are not available for all runs, and/or days. Fortunately, the weather conditions were very constant during the trials. Possibly, meteorological data from other days may be used in TARGAC. The effects of variations in meterological data, and of using data from other days, on the range predictions will have to be assessed.

Target and background
In TARGAC, there are two target parameters that affect the range predictions: effective target dimension, and thermal contrast between target and background (see 2.1.3). Since the BEST TWO targets are not in the TARGAC menu, the effective dimension of the selected target may differ from that of the target in the field. Further, neither the thermal contrast model in TARGAC (TCM2), nor the temperature measurements in the field can provide the inherent thermal contrast of the targets with high accuracy. First, the model calculates temperatures for the selected built-in target, and for a built-in background in TARGAC that does not exactly match the background in the test. Second, TCM2 only calculates *mean* target and background temperatures, whereas background temperature varies from location to location. In BEST TWO, temperature differences of more than 10 K were found for areas lying only several meters apart (Andersen, 1991). Third, field measurements of target and background temperatures may be used, but these were only

carried out at one location, and some time before a run. During a run, no measurements were carried out. Therefore, also measured inherent contrasts are inaccurate.
Thus, the effects of using a wrong effective target dimension and target and background temperatures on the range predictions have to be calculated.

MRTD
The MRTD was estimated in the field with an "objective method", using an apparatus that was under development (de Jong et al., 1991). This means that the accuracy of the MRTD may be limited. Bakker & Roos (1989) present a large number of objective MRTD measurements with this apparatus, together with subjective MRTD's. On the basis of their results, we estimate that with the objective method the error in cut-off frequency, for example, may be up to 20%.

4.2 Methods

The previous section shows that it is necessary to assess the sensitivity of the TARGAC output for variations in time and date, target type (effective dimension and temperature), background type (temperature) and MRTD. Two BEST TWO "standard situations" were defined as a reference for the sensitivity analysis. For the first, meteorological data were taken from an afternoon trial (July 27, 15:00 hrs). The second one corresponds to an early night trial (August 3, 23:00 hrs). An exercised T62 tank (TARGAC menu target no. 3) was selected as target, and the background was a grass field (menu background no. 17). The horizontal MRTD, which corresponds to the vertical resolution of the viewing device including the videorecorder, was chosen.
Recognition range predictions were always made at 5 probability levels: 90, 70, 50, 30 and 10% correct. Apart from the range predictions, also target and background temperature, as calculated by TCM2, will be considered.

4.3 Results of the sensitivity analysis

4.3.1 *The effect of time and date*

Fig. 1 presents the predicted detection and recognition ranges for the two standard situations. The predicted ranges are almost identical for afternoon and night. Predictions for other days, or other times of the day yield similar results. TCM2 predicts that inherent thermal contrast is high and is only slightly affected by the time of day: contrast is 9.2 K for the afternoon situation and 8.7 K at night. We conclude that, for the BEST TWO situation, the time of day or date are of minor importance with respect to predicted acquisition range.

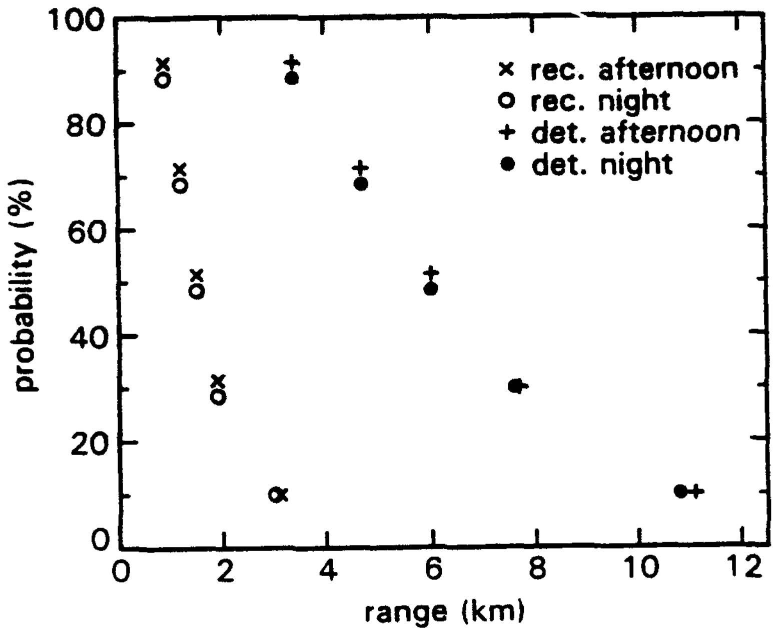


Fig. 1 TARGAC detection and recognition range predictions for the two BEST TWO standard situations at 5 levels of probability: 90, 70, 50, 30 and 10%. Predictions for afternoon and night are very similar. Where the values coincide, the symbols are shifted slightly in vertical direction. See text for details.

4.3.2 *The effect of acquisition level: transmission losses*

The effect of transmission losses through the atmosphere can be assessed by comparing the predicted ranges for recognition and detection in the following way. According to the Johnson criteria (on which the sensor performance model is based), the number of resolvable line pairs required for target *recognition* at a certain probability level, is four times the number of line pairs required for *detection* at the same probability level. For example, a 50% recognition probability requires 4 line pairs to be resolved across the target, whereas a detection probability of 50% requires only 1 line pair (see 2.1.3). This means that, if atmospheric effects are negligible, the ratio between detection and recognition ranges should be 4:1. Atmospheric effects will reduce this ratio.

Ratios between the detection and recognition ranges for the two standard situations (plotted in Fig. 1) are equal to 4.0 at all probability levels higher than or equal to 30%. This means that, for ranges below 7.0 km, transmission losses are negligible. The longest target range in the field trial was 4 km. Thus, for the BEST T VO situation, atmospheric effects do not play a role, at least if thermal contrast is high.

4.3.3 *The effect of target type*

TARGAC predictions were made for several target types. In order to change thermal contrast over a wide range, both targets in "exercised" and "off" states were chosen. The results are presented in Table 1. The first and second column give the target type and its effective, or minimum, dimension. The predicted inherent contrasts for the day and night situation are given in the third and fourth column, respectively. The two rightmost columns present the predicted 50% recognition ranges r_{50} for day and night.

Thermal contrast

Table 1 shows that thermal contrast varies over a wide range (-0.29 to 10.6 K). The largest differences occur between "off" and "exercised" targets. However, thermal contrast only has a small effect on predicted acquisition range: for each target, the four predicted acquisition ranges are very similar. Thus, a large influence of target temperature can only be expected for contrasts that are very close to zero (within tenths of degrees), which means that in most cases thermal contrast will not be the limiting factor for the predicted acquisition range.

TABLE 1: Effect of target type on TARGAC predictions.

target	target	inherent contrast (K)		r_{50} (km)	
type	height (m)	afternoon	night	afternoon	night
T62 (tank), off	2.2	2.6	1.5	1.5	1.2
T62, exercised		9.2	8.7	1.5	1.5
ZIL (Truck), off	2.6	3.3	-0.29	1.8	1.8
ZIL, exercised		5.8	2.2	1.8	1.7
T72 tank, off	2.3	2.9	3.6	1.6	1.5
T72, exercised		10.6	11.3	1.6	1.6
BRDM-2 (APC), off	2.1	2.0	1.3	1.4	1.3
BRDM-2, exercised		3.8	3.3	1.4	1.4

Target effective dimension

When atmospheric effects are negligible, predicted range is expected to be proportional to the effective dimension of the target. This is because a number of line pairs has to be resolved across the target effective dimension. For ground-to-ground target acquisition, the minimum dimension is target height. For the targets in Table 1, the average ratio between the predicted recognition range r_{50} (the two rightmost columns) and target height (the second column) is 0.67 +/- 0.04, which means that the expected proportionality is there. Thus, effective target dimension is a parameter that has a large influence on the model output.

4.3.4 *Background type*

The test field in Mourmelon mainly consisted of dry grass with bushes. However, due to the frequent use of the approach routes, bare soil came up and hot tracks appeared during the test, which were seen as white lines on the thermal imagery. At the longest ranges, the targets were seen against a background of wood. Several background types from the TARGAC menu were chosen to assess the possible influence of background type on the predicted ranges. The results are given in Table 2. Although the temperature is different for different backgrounds (contrast between target and background varies between 12.0 and 5.0 K), there is no effect of background type on recognition range.

TABLE 2: Effect of background type on TARGAC predictions

target type	target height (m)	r_{50} (km)
Leopard 2	2.50	1.8
AMX-30	2.30	1.6
PRI	2.60	1.8
PRAT	2.60	1.8
AMX-10	1.90	1.3
Truck	2.80	1.9
mean target	2.45	1.7

4.3.5 *MRTD*

When the apparent contrast is high, which is the case for the BEST TWO situation, the acquisition threshold will be determined in the high contrast - high spatial frequency region of the MRTD curve, and acquisition range may be expected to be *proportional to* the cut-off frequency of the MRTD-curve. Acquisition ranges were calculated for the two "standard situations", using both the horizontal and vertical MRTD-curves for the thermal imager (de Jong et al., 1991). The results show that he ratio between cut-off frequency and recognition range is indeed constant. Hence, the second parameter that has a large influence on the model output is the cut-off frequency of the MRTD, and an error in this frequency (which may be up to 20%, see 4.1) directly affects predicted acquisition range.

4.4 Verification of the results for other BEST TWO situations

The sensitivity analysis was carried out for the two "standard situations" only. In order to check whether the results of the sensitivity analysis also apply to all other BEST TWO situations, TARGAC predictions were made for all the BEST TWO runs that we had meteorological data for. As a tank, we used the exercised T72 tank; for the Truck we took the exercised ZIL. The BRDM-2 was used for the APC runs, and the grass field was chosen as background. The horizontal MRTD was chosen, which describes the vertical resolution of the system.

Predicted ranges for the same target under different conditions never differ by more than 0.1 km (which is the precision of the TARGAC output). Thus, the results of the sensitivity analysis apply to all conditions that we have meteorological data for.

4.5 Conclusions

The results of the sensitivity analysis for the BEST TWO situation can be summarized as follows:

1. Due to the excellent atmospheric conditions, predicted recognition range is almost independent of time or day.
2. Due to the excellent atmospheric conditions, recognition range is almost independent of target and background temperature. Therefore, it is not necessary to have an accurate estimate of target and background temperature. This is a very important result because large temperature differences were found between different locations in the field, and neither TCM2 nor the field measurements can provide the inherent contrast with high accuracy. For the calculations it makes no difference whether target and background temperatures calculated by TCM2 are used, or those measured during the trials.
3. There are only two input parameters that significantly influence the TARGAC outcome: predicted range is directly proportional to the effective dimension of the target, and to the cut-off frequency of the MRTD curve of the viewing device. The effective dimension of the BEST TWO targets is known with high accuracy (see 5.2). The error in the MRTD cut-off frequency may be up to 20%. Thus, a difference of up to 20% between actual and predicted ranges may be ascribed to inaccuracies in the values of the input parameters.
4. Because the influence of thermal contrast and atmosphere on recognition range are negligible, only the routine that describes electro-optical and human visual system performance, the NVESD Static Performance Model, can be tested with the BEST TWO observer data. This means that the results of the present evaluation may also be relevant for other models that are based on the Johnson approach.

5 TARGAC PREDICTIONS FOR THE BEST TWO RUNS

In the previous section it was shown that, due to the excellent atmospheric conditions during the BEST TWO trials, the predicted probability *vs.* range relationship for recognition depends significantly on only two parameters: effective target dimension and cut-off frequency of the MRTD of the viewing device. Here we will show that the TARGAC predictions for the entire set of BEST TWO runs can be described with a *single* equation that contains these two parameters. Such a simplified description of the predictions is convenient for two reasons:

First, the BEST TWO targets are not part of the TARGAC menu, which means that acquisition ranges for these targets have to be deduced from predictions for standard menu targets. With an equation that contains the effective target dimension, we can directly calculate recognition ranges for these targets.

Second, in a single run, TARGAC calculates ranges for only three probability levels (the improved version calculates five ranges, see 2.3). For the evaluation (see section 6), the entire probability *vs.* range relationship is required. Calculation of the entire curve would require a number of TARGAC runs for each condition. The equation we derive specifies the entire curve.

5.1 *Derivation of the probability vs. range equation*

In 4.3.3 and 4.3.5 it was shown that recognition range is directly proportional to target effective dimension and MRTD cut-off frequency. Since the Static Performance Model predicts a recognition probability of 50% if 4 line pairs can be resolved across the effective target dimension (2.1.3), it follows that:

$$r_{50} = \frac{f_{MRTD} \cdot D_{TARGET}}{4} \qquad (1)$$

where r_{50} is the recognition range (in km) at the 50% probability level, f_{MRTD} is the MRTD cut-off frequency (in lp/mrad), and D_{TARGET} is the effective dimension of the target (in m). Equation 1 is confirmed by the results of the calculations in section 4 (e.g. Table 1).

The probability versus range relationship can be described very well with an s-shaped curve which is known as the Weibull function. The relationship is given by:

$$P = \left[1 - 2^{-\left(\frac{r_{50}}{r}\right)^{s}} \right] \cdot 100\% \qquad (2)$$

where P is the predicted probability of a correct recognition, and r is the target range. The parameter s determines the steepness of the curve and is set to s = 2.32 for an optimal fit to the TARGAC predictions. Fig. 2 confirms that this function nicely coincides with the predictions for one of the "standard situations" (see 4.3).

Equation 1 and 2 can be combined to yield the following equation which describes the entire set of TARGAC recognition range predictions for the BEST TWO situation:

$$P_{\mathrm{TARGAC,BEST\ TWO}} = \left[1 - 2^{-\left(\frac{f_{\mathrm{MRTD}} \cdot D_{\mathrm{TARGET}}}{4\,r}\right)^{s}} \right] \cdot 100\% \tag{3}$$

or, inversely:

$$r_{\mathrm{TARGAC,BEST\ TWO}} = \frac{f_{\mathrm{MRTD}} \cdot D_{\mathrm{TARGET}}}{4} \cdot \left[- {}^{2}\log\left(1 - \frac{P}{100}\right) \right]^{-\frac{1}{s}} \tag{4}$$

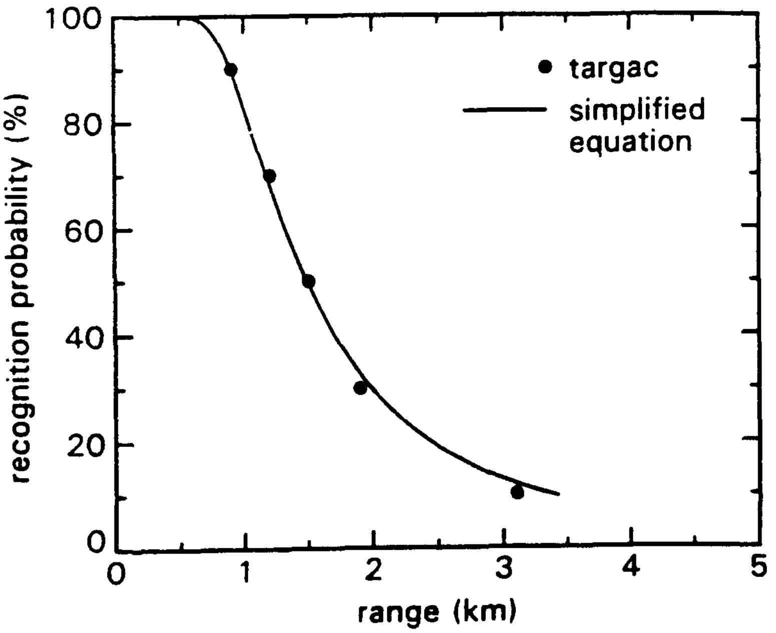


Fig. 2 Comparison of the results of the simplified equation (solid line) with the TARGAC predictions (filled circles) for the BEST TWO standard situation.

5.2 Range predictions for the BEST TWO targets

Equation 4 can be used to calculate the TARGAC range predictions for the specific targets that were used in the BEST TWO trials, if their effective dimension and the MRTD cut-off frequency of the viewing device are known. The results are presented in Table 3. The BEST TWO targets are listed in the leftmost column. The next columns give the estimate of their effective dimensions and the corresponding 50% recognition ranges (r_{50}), calculated with equation 4. The bottom row gives a 50% recognition range for a 'mean BEST TWO target', which will be used in one of the analyses in the next section. Note that the probability *vs.* range relationship is the same for each target (and each run), except for a single factor that is determined by the effective target dimension (equation 3).

TABLE 3: Target effective dimensions and predicted recognition ranges (at the 50% probability level) for the targets used in BEST TWO.

background type	temperature (K) (afternoon)	temperature (K) (night)	r_{50} (km) (afternoon)	r_{50} (km) (night)
deciduous trees	303.5	293.3	1.5	1.5
dirt road, dry	309.0	291.0	1.5	1.5
grass field	304.1	291.9	1.5	1.5
standard sand	304.6	291.1	1.5	1.5
foliage growing and sparse	300.6	290.7	1.5	1.5

5.3 Conclusions

TARGAC recognition range predictions for the entire set of BEST TWO runs can be described with a single equation, which contains only two input parameters: effective target dimension and MRTD cut-off frequency. Such an equation is convenient for the evaluation of the model since complete probability *vs.* range curves for the BEST TWO targets can be calculated at once.

It is also shown that, after our modifications in the software, the TARGAC calculations for the BEST TWO situation are in agreement with the Static Performance Model predictions, as may be expected.

6 EVALUATION OF RANGE PREDICTIONS

The evaluation of TARGAC will take place at various levels of complexity.
In section 6.1, the TARGAC predictions for the BEST TWO targets (see 5.2) will simply be plotted together with the observer data that are presented in Appendix 2. This gives a qualitative impression of the accuracy of the model predictions.
In section 6.2, the TARGAC predictions will be compared to the overall mean score of all runs. Originally, the Johnson criteria were based on mean acquisition performance over a large number of conditions. Therefore, it is useful to check their validity in this respect.
Finally, a complete quantitative comparison will be made between the set of individual datapoints and the TARGAC predictions. In the BEST TWO observer data, large differences in performance were found due to factors such as: target type, target distance, approach route and time of day. The sensitivity analysis showed that the TARGAC model predictions depend on only few of these factors. Thus, the model cannot account for at least part of the variation in the observer data. This part, which is called the "unexplained variance", will be determined in section 6.3. It is of obvious importance to know how large the unexplained variance is. It directly provides a measure of the reliability of the acquisition ranges predicted by the model. If the amount of unexplained variance is small, the model is able to make reliable predictions for individual cases or trials (for example, for a T62 tank on a grass field at 3000 m at 2:00 PM). If the amount of unexplained variance is large, the model is not applicable to individual cases and can only be used to predict overall mean performance.

6.1 Qualitative comparison for individual runs

The complete set of recognition performance data from the BEST TWO observer experiments, described in section 3.1, is presented in Figs. A9-A12 in Appendix 2, together with the corresponding TARGAC predictions for these runs. The set consists of 63 plots of recognition performance *vs*. target range. Each plot typically consists of 10-15 datapoints. Filled circles represent the averaged observer scores, and solid lines represent the TARGAC predictions. The standard error of the mean of the observer scores, s_P, was smaller than 10% (see Appendix 3).

Three typical examples from this set are given in Fig. 3. In Fig. 3a, the prediction is reasonable: both measured and predicted recognition performance gradually decrease with target range. However, TARGAC underestimates observer performance. In Fig. 3b, predicted recognition performance is far too low. The data show that target recognition probability is better than 80% for ranges up to 4000 m, whereas the predicted probability is below 80% at a range of 1000 m. At a distance of 4000 m, TARGAC predicts a recognition probability of less than 10%! Fig. 3c shows that recognition performance does not simply decrease with target range, but changes rapidly with the exact position of the target. The recognition probability is high at distances below 1600 m, at 2200 and 2300 m, and at 3900 m. At intermediate distances, e.g. at 1900 m and 2600 m, recognition probability is very low. This behaviour was termed "target-terrain interaction" (see Chapter 4), and it cannot be described satisfactorily with a monotonously decreasing function. On average, predicted performance is again far too low.

In general, the following conclusions can be drawn:

- The predicted curves fall below most of the datapoints. This means that, on average, the TARGAC predictions are conservative: at a certain range the predicted probability is too low, or, equivalently, the predicted ranges for a certain probability level are too small. For individual points the deviation can be considerable: the data often show a high recognition probability at distances near 4000 m, whereas TARGAC predicts a probability less than 10% for these ranges.
- The monotonously decreasing performance curve that TARGAC predicts, is found only in a number of cases.
- In some cases, strong undulations in the relation between target distance and acquisition performance are found, because not only target range, but also the local conditions are an important determinant of observer performance. Strictly, a probability *vs.* range relationship does not exist in these cases. Such behaviour is not predicted by the model.
- TARGAC is not suitable for predicting performance for individual targets.

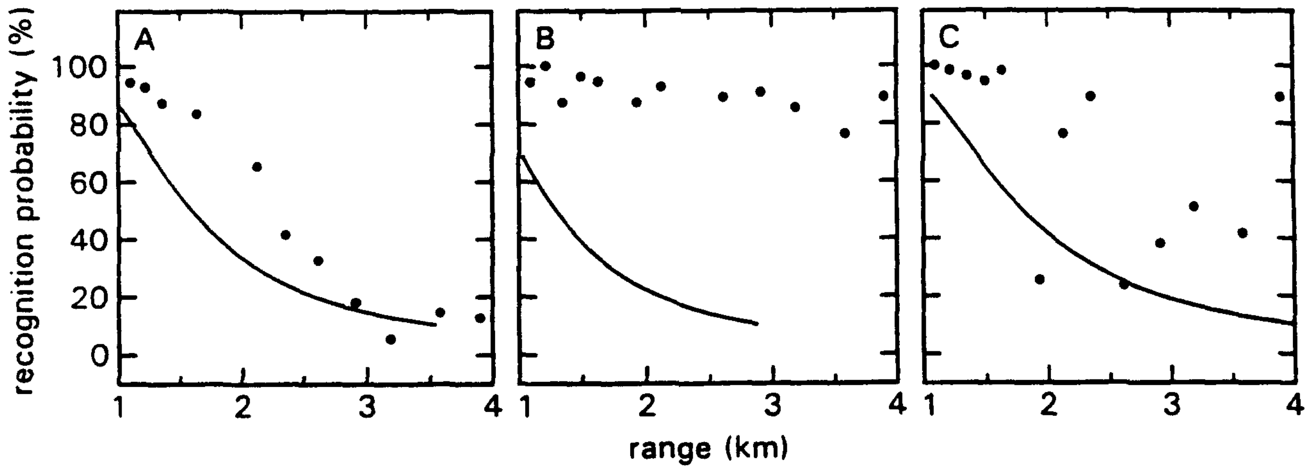


Fig. 3 Comparison of observer performance *vs.* target range and the corresponding TARGAC predictions for three typical examples. Filled circles: observer recognition scores. Solid lines: TARGAC predictions. A: measured and predicted performance gradually decrease with target range. TARGAC underestimates observer performance. B: predicted recognition performance is far too low at all ranges. C: TARGAC is unable to predict the large undulations in the observer scores.

6.2 Comparison with overall mean observer performance

Since TARGAC does not appear to be suitable to predict performance for individual targets, its predictions will now be compared with the overall mean performance of observers over a large number of targets and runs. This seems to be a reasonable approach because originally the Johnson criteria were also based on mean acquisition performance over a large number of conditions.

Data sets A, B, and C, defined in section 3.1, will be used, and the comparison with the TARGAC prediction for the 'mean' BEST TWO target (see Table 3) is shown in Fig. 4. Using the prediction for a 'mean' target makes sense because all targets in the set were presented to the observers roughly equally often, and predicted ranges for the six targets do not differ considerably. Filled circles represent mean observer performance (averaged per distance). The solid line indicates the TARGAC predictions for the BEST TWO mean target. The dashed line represents the best fit of equation 2 (see 5.1) to the observer data and will be shown to correspond to TARGAC range predictions increased by a factor of 1.80 (see 6.3).

Data set A is the largest set. Target images were presented in the "pop-up" presentation order. Data sets B and C (a subset of A) correspond to the same set of images, but for data set B the images were presented as an ordered sequence, simulating a target approach.
Fig. 4 shows that:

- TARGAC indeed underestimates mean observer performance for all data sets.
- The mean recognition probability for the observers is never below about 50% correct, even at the longest target range (4 km). This means that the shape of the measured probability *vs.* range relationship is not known for lower probabilities.
- For data sets A and B, the overall mean probability decreases with target range, which was not the case for the scores for individual runs (Fig. 3). This is because averaging over many runs diminishes the effects of target-terrain interactions. The results for set C show more residual terrain interactions because it is a small subset of A.
- Comparison of data sets B and C shows that for "approaching" targets (B) the data conform much better to a monotonously decreasing function than for "pop-up" targets (C). This is because the accumulation of information during a target approach helps to reduce the effects of the target-terrain interactions , see Chapter 4.
- For the "approaching" presentation order (data set B), the slope of the dashed curve fits nicely to the data. For "pop-up" targets, the predicted curve is too steep. A shallower function would fit better. This difference in slope is most pronounced for data set A. This might suggest that the model should be made to accommodate these different target behaviours.

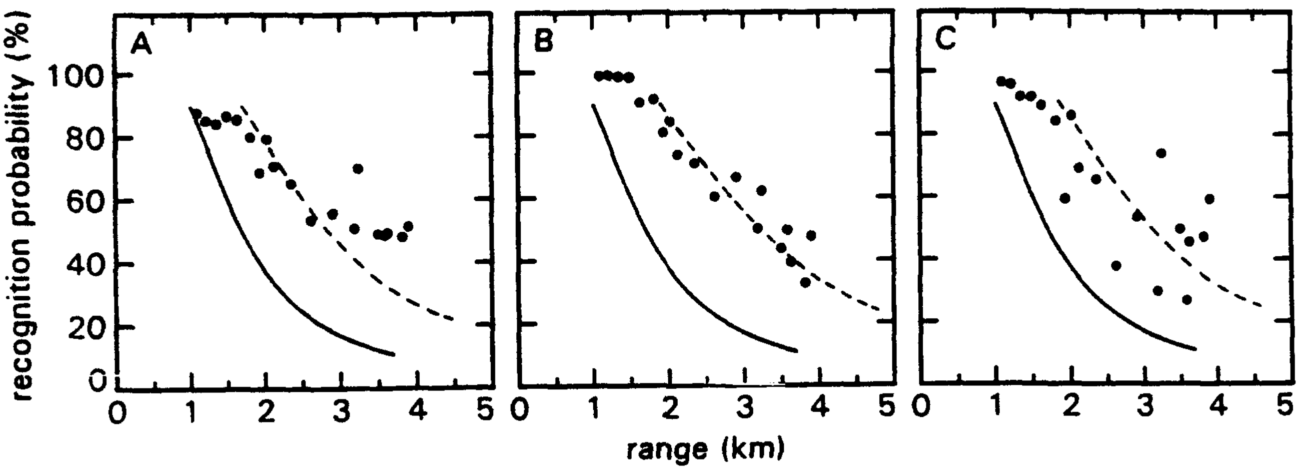


Fig. 4 Comparison of overall mean observer performance for data sets A, B and C, and TARGAC predictions. Filled circles: mean observer recognition scores. Solid lines: TARGAC predictions. Dashed line: best fit of equation 2 (see 5.1) to the data. TARGAC predictions are far too conservative. See text for details.

6.3 Comparison with observer performance for individual trials

In the previous sections, it was shown that overall *mean* recognition performance can be described reasonably well as a monotonously decreasing function of target range. The fit of the model to the mean observer data could be much improved by applying a single correction factor and possibly a small change in the steepness s (equation (3) or (4)). Apart from the *mean* performance, it is worthwhile to know how well TARGAC predicts the performance for *individual trials*. The previous sections have shown that there is much variation in the observer data, but the model predicts only a single curve, with an unknown confidence interval.

In order to determine the unexplained variance in the observer data, these will be regarded as a large set of single (probability *vs.* range) points, and a point by point comparison will be made between actual target range, and the range that is predicted by the model (range comparison). This analysis will yield a distribution which gives the unexplained variance in the data due to all parameters that were varied in the experiments (including the effects of local conditions). The variance provides an indication of the quality of the model predictions for individual trials.

6.3.1 *Procedure*

Each datapoint represents a probability of correct recognition *P* for a target at range *r*. At this probability level, the model predicts a target range *r'*. For each datapoint in the set, a ratio *r/r'* between actual and predicted range is calculated. If the model makes a correct prediction, the ratio $r/r' = 1$. If the predicted range is too short, the ratio will be larger than 1, if it is too large, the ratio is smaller than 1. Fig. 5 gives an example of the procedure for a few datapoints. It is convenient to transform the values to a log scale, because correct predictions are centered at 0, and an over- or underestimate of the range by the same factor (for example the predicted range is twice or half the actual range) are equally shifted in opposite directions along the axis.

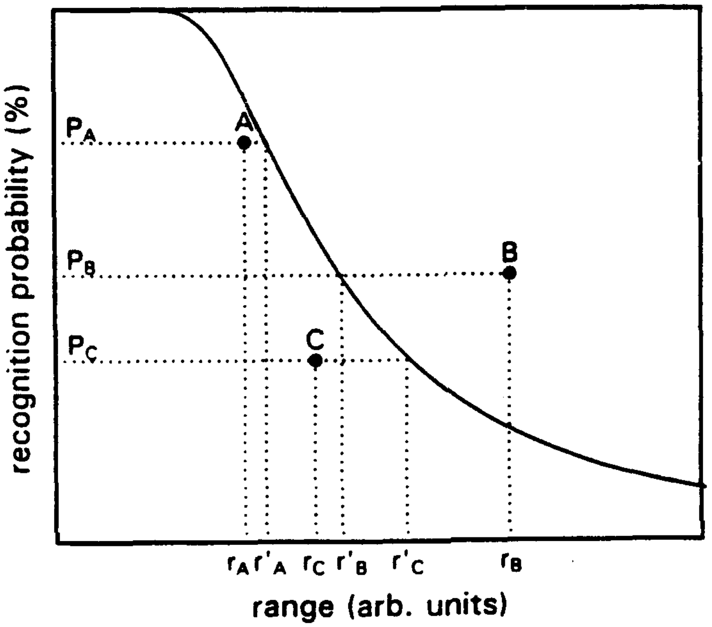


Fig. 5 Example of the point by point comparison between measured and predicted recognition performance. The solid line represents the model prediction, A, B and C are datapoints. For each datapoint, probability P corresponds to an actual target range r and a predicted target range r'. Point A: predicted and measured range are almost identical. The ratio $r/r' \approx 1$. Point B: actual range is longer than predicted range at the same probability level: $r/r' \approx 1.5$. Point C: actual range is much smaller than predicted range at the same probability level: $r/r' \approx 0.75$.

A set of datapoints gives a (dimensionless) distribution of log (r/r') - values. An example of a distribution is given in Fig. 6. Mean and variance of the distribution directly provide a measure of the accuracy of the model. The *mean* of the distribution, «log (r/r')», indicates how well the model predicts overall mean performance. If «log (r/r')» = 0, overall mean performance is correctly predicted. A shift of the distribution along the log (r/r') axis means that, on average, predicted acquisition range is too long or too short. In that case, «log (r/r')» provides a range correction factor that will make the model predict overall mean performance correctly. The *variance* $\sigma^2_{(r/r')}$ in the distribution indicates how well the model predicts acquisition performance for individual trials. Part of the variance, $\sigma^2_{(r/r'),obs}$ is due to statistical errors in the observer scores, because an error in the recognition probability P leads to an error in *r'*, and hence in log (*r*/*r'*). In Appendix 3 it is shown how $\sigma^2_{(r/r'),obs}$ is calculated from the standard error σ_P in *P* . The remainder of the variance can be ascribed to incorrect predictions by the model. Thus, if the evaluation yields that $\sigma^2_{(r/r')} \approx \sigma^2_{(r/r'),obs}$ the model will be accepted because it correctly predicts observer performance (within the accuracy of the measurements) as a function of the parameters that were varied in the experiment (e.g. target type, part of day). On the other hand, if $\sigma^2_{(r/r')} \gg \sigma^2_{(r/r'),obs}$ the unexplained variance is mainly due to differences between the model predictions and actual observer performance. In that case, the 95% confidence interval of the distribution, [«log (*r*/*r'*)» - $2\sigma_{(r/r')}$, «log (*r*/*r'*)» + $2\sigma_{(r/r')}$], indicates the quality of the model predictions. A wide uncertainty interval means that the model is not able to make reliable predictions for individual trials.

An elegant property of the procedure is that the entire set of datapoints, irrespective of the run or the circumstances under which they were collected, can be analyzed at the same time, yielding a single distribution. The analysis can also be carried out for subsets of the data. In section 7 it will be shown that analysis of subsets may be used to discover which factors significantly contribute to the variance in the distribution. If these factors are known, the model may be improved.

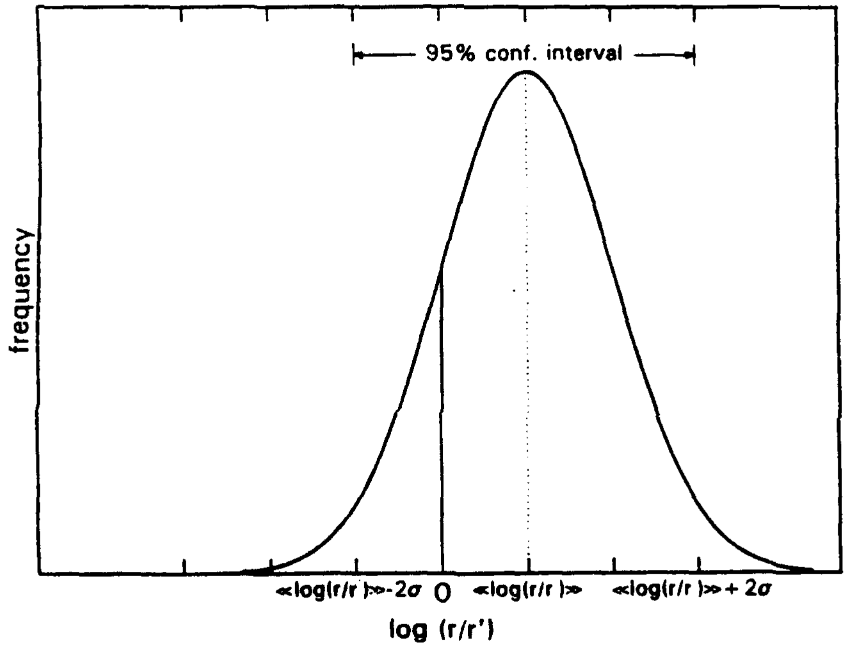


Fig. 6 Example of a distribution of *r*/*r'* - values on a log-scale. If the mean of the distribution, «log (*r*/*r'*)», is equal to 0, the model correctly predicts overall *mean* performance. A shift of the distribution means that, on average, predicted acquisition range is too long or too short (in the example, predicted ranges are too short). The variance σ^2 in the distribution indicates how well the model predicts acquisition performance for *individual* trials (see text).

6.3.2 *Results*

Fig. 7 shows the results of the evaluation for data set A (38 runs, "pop-up" presentation order). In Fig. 7a, the ratio between actual and predicted range (*r/r'*) is plotted for individual data points. Only datapoints with a recognition probability between 20% and 80% are considered (230 points). This is because at very high or very low probabilities, (*r/r'*) may take unrealistic values (see section 6.3.3). The standard deviation $\sigma_{(r/r'),obs}$ in the (*r/r'*) values, due to the statistical error in the observer scores, is presented in the upper right-hand corner of the figure. In Appendix 3 it is shown that $\sigma_{(r/r'),obs}$ = 0.03 - 0.06 log units for probabilities between 20% and 80%. Fig. 7b shows the histogram of the distribution of log (*r/r'*)-values, based on the datapoints in Fig. 7a. Note that this is a roughly normal distribution. The mean and the 95% confidence interval of the distribution are indicated in Fig. 7a by the fat dashed line and the two dotted lines, respectively.

It is clear, that the quality of the model predictions for individual trials is not very good. First, the mean of the distribution is 0.23 which means that, on average, actual ranges are $10^{0.23}$ = 1.70 times the predicted ranges. For the whole study, the average range correction factor is 1.8. This difference cannot be ascribed to inaccuracies in the values of the input parameters (see 4.5).
Second, the standard deviation $\sigma_{(r/r')}$ = 0.17 on a log scale, which is much larger than $\sigma_{(r/r'),obs}$. The 95% confidence interval is [-0.11,. 0.57]. On a linear scale, this interval is [0.8, 3.7], spanning a factor of almost 5! In words, there is a 95% probability that the actual recognition range for an individual trial falls between 0.8 and 3.7 times the range that is predicted by TARGAC. Even after correction for the overall mean acquisition range, the actual range may be more than twice, or less than half, the predicted range. Thus, the model is not very good at predicting how observer performance depends on the prevailing conditions in the field.

For data set B (15 runs, "approaching" presentation order) and C (15 runs, "pop-up" presentation order), similar results are found. The mean shifts between the data and the TARGAC predictions are 0.28 for set B (70 data points) and 0.27 for set C (65 datapoints) on a log scale, which correspond to an underestimate of the actual range by a factor 1.90. The uncertainty interval is very large in both cases, spanning a factor of about 4 for the "pop-up" presentation order (data set C) and 3.3 for the "approaching" presentation order (data set B). The smaller interval for the latter case is is found because the effects of target-terrain interactions are less pronounced using this presentation order.

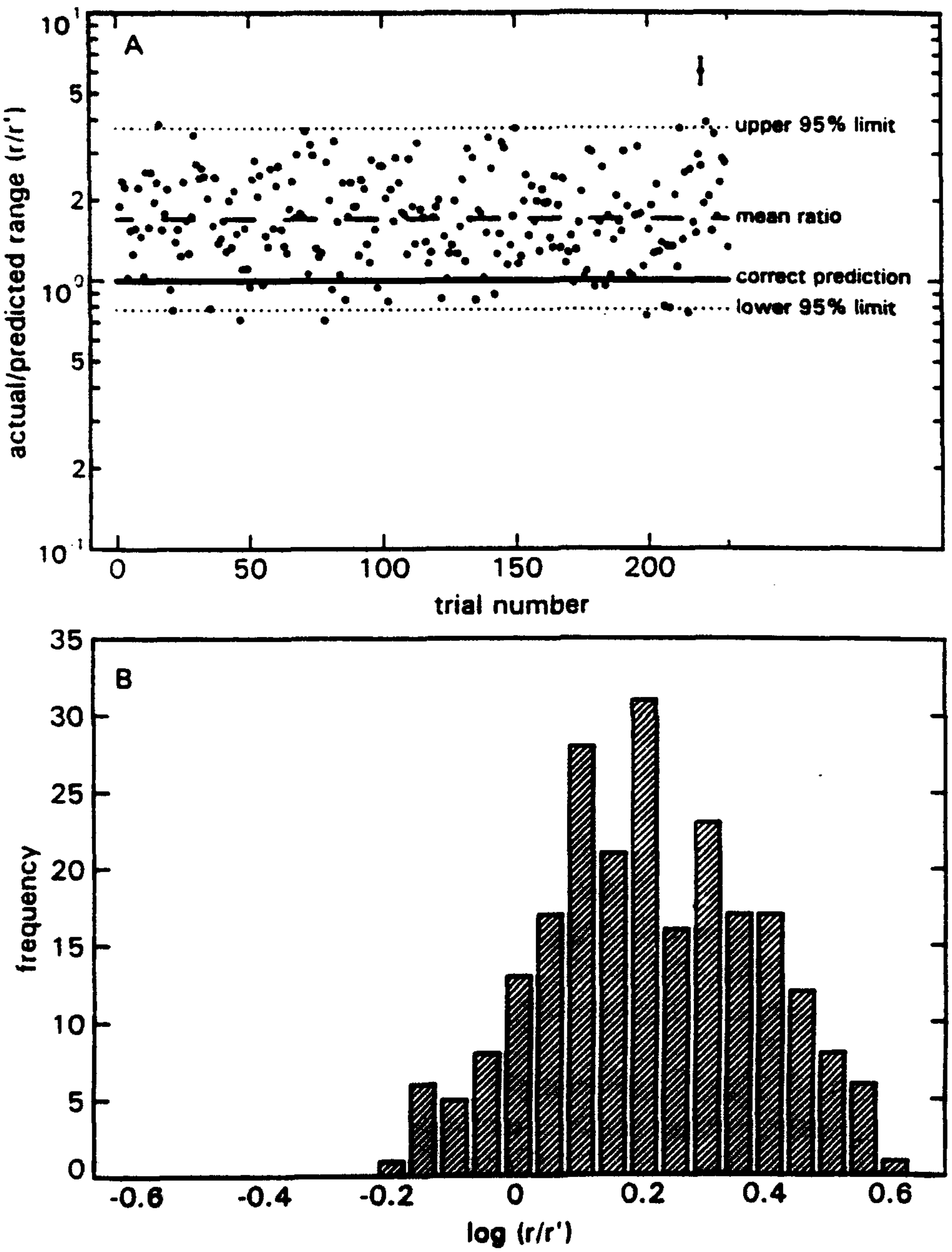


Fig. 7 Comparison between measured and predicted recognition performance for individual trials. A: Ratio between actual and predicted range (r/r'). With a perfect model, the points would be concentrated around the solid line ($r/r' = 1$). The dashed line corresponds to the mean of the log (r/r') distribution. The dotted lines indicate the boundaries of the 95%-confidence interval. The error in observer scores is shown in the upper right-hand corner of the figure. B: Histogram of the log(r/r') distribution. Mean is 0.23; The standard deviation is 0.17. It is clear that the model is not very good at predicting recognition performance for individual trials.

6.3.3 *High and low probability levels*

At very high or low probabilities, small errors in the observer scores lead to large errors in the value of *r/r'*, because the slope of the predicted probability *vs.* range curve is very shallow in those regions. In Appendix 3, it is shown that for the analysis it is safe to use only observer data with a probability level between 20% and 80%. Using data with a higher or lower probability level may lead to a broadening of the log (*r/r'*) distribution. The effect of using a larger interval on the distribution was estimated for the three data sets A, B and C. The effect was similar for all sets. Both mean and standard deviation of the log (*r/r'*)-distribution are only slightly increased (approximately by 0.02 and 0.01 log units, respectively) when the interval is changed from 20-80% to 10-90%. A further increase is not possible since TARGAC predictions are only defined between 10% and 90%. As a conclusion, the extent of the interval does not have a very large effect on the results.

6.4 Conclusions

There are important differences between recognition performance, as measured in observer experiments, and the TARGAC predictions for the BEST TWO situation:

1. The model predictions are too conservative. On average, TARGAC underestimates recognition range by a factor 1.8. This ratio is similar for "pop-up" and "approaching" targets. The difference cannot be ascribed to inaccuracies in the values of the model input parameters.

2. TARGAC does not make accurate predictions for individual trials. The analysis shows that the 95%-confidence interval, is roughly given by 0.9-3.6 times the predicted acquisition range, thus spanning a range of a factor 4.

7 POSSIBLE SOURCES OF THE UNEXPLAINED VARIANCE

In the previous section it was shown that, when TARGAC predictions are compared with actual observer performance for individual trials, there is a large amount of unexplained variance. The variance cannot be ascribed to the statistical error in the observer scores. The conclusion is that the model does not predict how observer performance depends on field factors or parameters that were varied in the experiment.

If we can determine which factors contribute significantly to the unexplained variance, it may be possible to make a model that better predicts performance for individual situations. Such factors can be found using Analysis of Variance. Suppose that, for

example, the effect of target type is not modelled correctly. In that case, the broad distribution of log (*r*/*r'*)-values that was found for entire dataset, is in fact a composition of narrower distributions with different mean shifts for different target types. The model could then be improved (i.e. the amount of unexplained variance would be reduced) by remodelling the effect of target type.
In a similar way, possible effects of other factors may be tested. Hypotheses that can straighforwardly be tested with the present data set, are:

a Target type has an effect on acquisition range, but the effect is not modeled adequately by simply taking its minimum dimension.
b Time of day has an effect on observer performance, although the model does not predict any differences.
c Part of the variance that was found in data set A, may be due to using both stationary and (head-on) moving targets. The model is only designed for stationary targets.
d Target-terrain interaction, as mentioned in section 6.1, has a considerable impact on acquisition performance. This hypothesis can for example be verified by comparing the results for different approach routes.

Analysis of Variance was used to determ:ne which of the above-mentioned factors have a *significant effect* on the log (*r*/*r'*)-distribution for the comparison between the TARGAC predictions and the observer data from set A (which is the largest data set). Only main effects are considered: interactions are not considered relevant for a first investigation.
The analysis shows that there are statistically significant effects (P < 0.05) of target type, time of day, and approach route on variance. No significant effect was found for (head-on) target motion. The effects are:

Target type
Mean range shifts («log (*r*/*r'*)») for most targets are quite similar to the mean shift found for the complete set (ranges do not differ by more than 20%). However, for one target (AMX-10) the shift is considerably larger. Predicted range for this target is much shorter than for the other targets because of its small height (see Table 3), but measured performance is in fact even slightly better than for the other targets. When optimal shifts are applied for each target type separately, the standard deviation of the distribution is reduced from $\sigma = 0.17$ to $\sigma = 0.15$, which means that most of the variance remains unexplained. Thus, the model cannot be improved considerably by remodeling the effect of target type.

Time of Day
Runs were divided into three categories: morning, afternoon and early night runs (in accordance with the division that was made during the BEST TWO trials, Valeton & Rogge, 1992, Reichart, 1993). There is a small but significant difference in mean range

shift for morning on the one hand, and afternoon and early night on the other. There is no relevant reduction of the width of the distribution when the effect of time of day is taken into account.

Approach route; target-terrain interactions
Mean recognition ranges for the Right approach route were approximately 20% longer than for the Left route. These routes were very close to each other, heading in approximately the same direction, and there exists no theoretical explanation why acquisition of targets on the Right approach route would be easier. This means that local factors are very important. Furthermore, the results of the observer experiments also show that for a single approach route, in general there is no monotonous relationship between recognition performance and target distance. These effects were ascribed to a strong interaction between target signature and local background ("target-terrain interactions", see Chapter 4). Thus, a large amount of variance may be ascribed to target-terrain interactions.

A model that takes into account the effects of target-terrain interaction, may become very complicated. Apart from mean or area contrast between target and background (which is modeled in TARGAC), there are many local factors that may influence acquisition performance, such as: edge contrast, internal target contrast, differences in target and background structure, or (small variations in) target orientation. Their effect on acquisition performance is unknown. By analyzing the BEST TWO images, it might be possible to determine the effect of some of the above mentioned local factors on recognition performance. Such an analysis goes beyond the scope of this report.

In conclusion, there seems to be no simple modification that would lead to a model that better predicts acquisition performance for individual trials, i.e. a modification that would considerably diminish the amount of unexplained variance that is found when TARGAC predictions are compared with actual observer performance.

8 DISCUSSION

TARGAC is a very comprehensive target acquisition model. The model combines modules for target and background characterization, atmospheric transmittance, and system/observer performance. These properties make TARGAC in principle to a very useful tool for military purposes, especially as a tactical decision aid (TDA). The evaluation of TARGAC, however, shows that the model does not predict recognition performance very accurately. The main differences between observed recognition performance and the model predictions for the BEST TWO situation, are:

1. TARGAC under-estimates the mean recognition range for the BEST TWO situation by a factor of about 1.8.
2. TARGAC predictions for individual cases have an uncertainty interval of roughly 0.9-3.6 times the predicted acquisition range, thus spanning a range of a factor 4.

The sensitivity analysis (section 4) showed, that for the BEST TWO situation only the module that describes electro-optical and human visual system performance, is responsible for the range predictions. Due to the excellent conditions, the effects of the outcome of the target-background contrast module and the atmospheric transmittance module on the range predictions are negligible. The system performance module in TARGAC is theoretically equivalent to 1-D ACQUIRE90, the one-dimensional option of the 1990 version of the NVESD Static Performance Model (the 2-D option of ACQUIRE is discussed below). A comparison between the TARGAC and the 1-D ACQUIRE90 recognition performance predictions for the BEST TWO standard situation (see 4.2) is presented in Fig. 8. Filled circles represent the TARGAC predictions, and open circles the corresponding 1-D ACQUIRE90 predictions. In ACQUIRE90, atmospheric transmission was set to 1.0, and target background contrast was set to 5.0 K. Varying the contrast had little effect on the ACQUIRE90 predictions. Evidently, the predictions made with the two models are identical.

The conclusion is that the results of the evaluation are not only relevant for TARGAC but for all models that are based on the 1-D NVESD Static Performance Model.

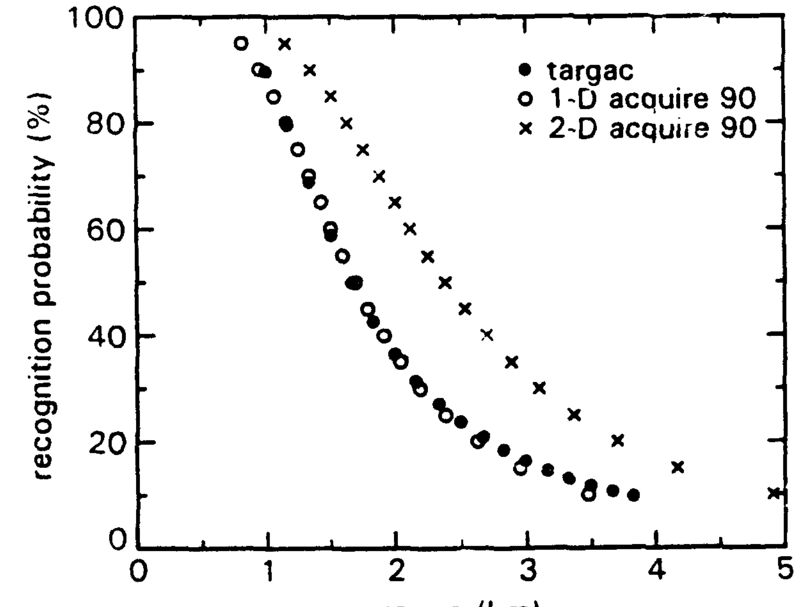


Fig. 8. TARGAC and 1-D ACQUIRE recognition performance predictions for the BEST TWO situation are identical. 2-D ACQUIRE predicts longer ranges.

Mean acquisition performance
As was shown in section 4, inaccuracies in the data that are input to the model may lead to an error of up to 20% in predicted range. The performance of the observers in the experiment may be relatively high, because they were trained on six targets in a specific situation (a single wheather condition, one terrain). Their score might have been lower if there were more uncertainties in their task. However, these factors are not large enough to explain the considerable difference between actual and predicted *mean* acquisition performance.

TARGAC may easily be adapted to correctly predict mean acquisition performance for the BEST TWO situation by changing the recognition criteria in the system performance module. The present version of the model predicts a recognition probability of 50% if four line pairs can be resolved across the effective dimension of the target. According to Ratches (1976), this is a conservative criterion (see 2.1.3). However, even the optimistic criterion that Ratches gives (a three line pair criterion, which would predict 4/3 longer ranges) is still too conservative. To correctly predict mean recognition range for the BEST TWO situation, a recognition probability of 50% should correspond to the resolution of 2.2 line pairs across the effective dimension of the target. In addition, it should be noted that the targets in BEST TWO were always in front view. It is well known that targets in side view are more easily recognized than targets in front view, but the 1-D Static Performance Model predicts equal ranges, because the effective dimension is target height in both cases. This means that, for targets in side view, prediction of mean recognition performance by TARGAC may be even further off.

Recently, a new version of the Static Performance Model (2-D ACQUIRE90) has been developed, that makes its predictions on the basis of two target dimensions and both the horizontal and vertical MRTD. In Fig. 8 it is shown, that the new version (crosses) predicts longer ranges than the 1-D model. It also predicts better acquisition performance for targets in side view than for targets in front view. For the BEST TWO situation, the 2-D version underestimates mean acquisition range by less than 30%, an error which is near the accuracy of the MRTD cut-off frequency. Thus, TARGAC may be improved by implementing the two-dimensional version of the Static Performance Model.

Acquisition performance for individual cases
The TARGAC user interface suggests that predictions can be made for *individual* targets under given conditions. For example, the model predicts the 50% recognition range for a T62 tank on a grass field at 2:00 PM. The evaluation shows that the model can not make such predictions. A very large uncertainty interval is found when the predictions are compared with data from a single observer performance experiment, using one target set, in one terrain, with one wheather condition, and one camera. The interval may be even larger if more conditions are investigated. The large amount of

unexplained variance is not due to possible errors in the input data, such as the MRTD-curve, because these only affect the mean of the confidence interval, and not its width. Nor is the width of the interval decreased if the original system performance model is replaced by the 2-D version. Apparently, the model is not sophisticated enough to deal with individual cases: factors that play an important role in target acquisition, are not modelled, or are not modelled correctly.

A model that better predicts acquisition performance for individual trials may have to be very complex. In section 7, it was shown that local factors play an important role in target acquisition. Thus, a correct prediction probably requires a detailed model of the effects of local factors on acquisition performance, and, of course, a detailed description of the local conditions as input to the model. Such a detailed modeling, if possible at all, may not be useful for practical purposes. The 'equivalent disc' model (van Meeteren, 1990) is based on a different approach: this model predicts acquisition performance for a *set* of targets, rather than predicting a range for each target separately.

It is likely that the width of the uncertainty interval depends on the type of terrain. For the BEST TWO trials, local factors were very important because of the inhomogeneity of the background, and consequently the uncertainty interval is large. If the background is more uniform, one expects that acquisition performance mainly depends on target distance (as a model predicts), resulting in a smaller 95% confidence interval. The dependence oî the reliability of the acquisition range predictions on terrain type, or on statistical information about the terrain, may be a topic of future research.

9 CONCLUSIONS AND RECOMMENDATIONS

The target acquisition model TARGAC was evaluated using BEST TWO observer performance data for recognition of targets in front view. Due to the excellent atmospheric conditions during the BEST TWO trials, recognition performance predictions are determined solely by the system performance module of TARGAC, which is equivalent to the 1-dimensional NVESD Static Performance Model. The main results of the evaluation are:

1. The model predictions are too conservative. On average, TARGAC underestimates recognition range by a factor 1.8. This ratio is similar for "pop-up" and "approaching" targets.
2. TARGAC does not make accurate predictions for individual cases. The analysis shows that the uncertainty interval roughly ranges from 0.9 to 3.6 times the predicted acquisition range, thus spanning a range of a factor 4.

3. The TARGAC predictions for overall mean performance can be improved by incorporating the 2-dimensional version of the Static Performance Model.
4. It is proposed that TARGAC predictions are not only presented as single numbers for acquisition probability *vs.* target range, but that some indication is given of the accuracy of the results, preferably in the form of a 95% confidence interval.
5. The version of TARGAC that was tested (PC version released in 1992) contained a number of software errors and minor problems. A number of corrections are suggested. Additional work in modularization and streamlining the model is recommended. It is also recommended that the model is given a more consistent and user-friendly user interface.
6. TARGAC and other models that incorporate the NVESD Static Performance Model sho"ld only be used to provide an indication of the actual acquisition performance.

ACKNOWLEDGEMENTS

We thank J. Varkevisser, A. Everts and A.J.C de Reus for expert technical assistance with the development of hardware and software for the observer performance experiments, and with the TARGAC validation analyses. We thank Dr. Patti Gillespie from ARL-BED for help with compiling the TARGAC input files, guidance and scientific support.

BIBLIOGRAPHY

Valeton, J.M., Bijl, P, & De Reus, A.J.C. (1992). The BEST TWO field trial: timetables of events, maps, and target distance data. Rep. No. IZF 1992 A-10, TNO Human Factors Research Institute, Soesterberg, The Netherlands.

Valeton, J.M., Rogge, J. (1992). The BEST TWO scenarios: an overview. Rep. No. IZF 1992 A-32, TNO Human Factors Research Institute, Soesterberg, The Netherlands.

Valeton, J.M., Bijl, P. (1992). Observer experiments with BEST TWO thermal images. Part 1: Design, Training and Observer selection. Rep. No. IZF 1992 A-33, TNO Human Factors Research Institute, Soesterberg, The Netherlands. (NATO Confidential).

Bijl, P., Valeton, J.M. (1992a). Observer experiments with BEST TWO thermal images. Part 2: Terrain interaction and Target motion. Rep. No. IZF 1992 A-34, TNO Human Factors Research Institute, Soesterberg, The Netherlands. (NATO Confidential).

Bijl, P., Valeton, J.M. (1992b). Observer experiments with BEST TWO thermal images. Part 3: Reliability of observer responses. Rep. No. IZF 1992 A-35, TNO Human Factors Research Institute, Soesterberg, The Netherlands. (NATO Confidential).

Bijl, P., Valeton, J.M. (1992c). Observer experiments with BEST TWO thermal images. Part 4: complete set of observer performance data. Rep. No. IZF 1992 A-43, TNO Human Factors Research Institute, Soesterberg, The Netherlands. (NATO Confidential).

Bijl, P., Valeton, J.M. (1994). Evaluation of Target Acquisition Models TARGAC using 'BEST TWO' observer performance data. Rep. No. IZF 1994 A-??, TNO Human Factors Research Institute, Soesterberg, The Netherlands. (In press)

REFERENCES

Andersen, E. (1991) BEST TWO Infrared signatures of the vehicles participating in the trial, Report No. DDRE N-7/1991, Danish Defence Research Est. Copenhagen, Denmark (NATO Confidential).

Bakker, S.J.M. & Roos, M.J.J. (1989) MRTD-metingen aan warmtebeeldcamera's (in Dutch). Rep. No. FEL-1989-118, TNO Physics and Electronics Laboratory, The Hague, The Netherlands (NATO Confidential).

Bartleson, C.J. & Grum, F. (1984). Optical Radiation Measurements, Vol. 5: Visual Measurement. Academic Press.

Danielian, F. (1992). An overview of the RSG 15 test. Proceedings of the Second Symposium on Measuring and Modelling the Battlefield Environment, p. 1, Paris, NATO.

Gillespie, P. (1991) TARGAC user guide. U.S. Army Research Laboratory, Battlefield Effects Directorate, White Sands Missile Range, NM, USA.

Gillespie, P. (1993). Personal communication.

Johnson, J. (1958). Analysis of image forming systems. Image Intensifier Symposium, Warfare Vision Branch, Electrical Engineering Dept., U.S. Army Engineer Research and Development Laboratories, Fort Belvoir, VA, 249-273.

Jong, A.N. de, Janssen, Y.H.L, Roos, M.J.J. & Kemp, R.A.W. (1991) Infrared and Electro-Optical experiments during Best Two by Research Group Infrared. Report No. FEL-91-A252, TNO Physics and Electronics Laboratory, The Hague, The Netherlands (NATO Restricted).

Meeteren, A. van (1990) Characterization of task performance with viewing instruments. J. Opt. Soc. Am. 7, 10, 2016-2023.

Ratches, J.A. (1976). Static Performance Model for Thermal Imaging Systems. Opt. Eng. 15, 6, 525-530.

Ratches, J.A., Lawson, W.R., Shields, F.J., Hoover, C.W., Obert, L.P., Rodak, S.P. & Sola, M.C. (1981). Status of Sensor Performance Modelling at NV&EOL. Night Vision & Electro-Optics Laboratory, Fort Belvoir, VA 22060, USA.

Reichart A. (Ed.) (1993) Second Symposium on Measuring and Modelling the Battlefield Environment, Volume 1A (unclassified). Technical Proceedings AC/243 (Panel 4) TP/1. SEFT/CTE/OSA.

Smith, R. & Corbin, T. (1992). Meteorological conditions and data for the BEST TWO field trial. Proceedings of the Second Symposium on Measuring and Modelling the Battlefield Environment, p. 3, Paris, NATO.

Vonhof, D.M. & Goessen, A. (1992). The effect of white phosphorous smoke and artillery dust clouds on target acquisition. Proceedings of the Second Symposium on Measuring and Modelling the Battlefield Environment, p. 19, Paris, NATO.

Vonhof, D.M. & Rogge, J. (1992). Reliability of observer responses using inventory thermal imagers. Proceedings of the Second Symposium on Measuring and Modelling the Battlefield Environment, p 23, Paris, NATO.

Wagenaars, W.M. (1990). Localization of sound in a room with reflecting walls. J. Audio Eng. Soc. 38(3), 99-110.

Wester, H.G. & Van de Mortel, F.W.M. (1990). Waarnemingsexperiment BEST TWO (in Dutch). Report, Royal Military Academy, Breda, The Netherlands.

APPENDIX 1: THE COMPLETE SET OF OBSERVER PERFORMANCE DATA

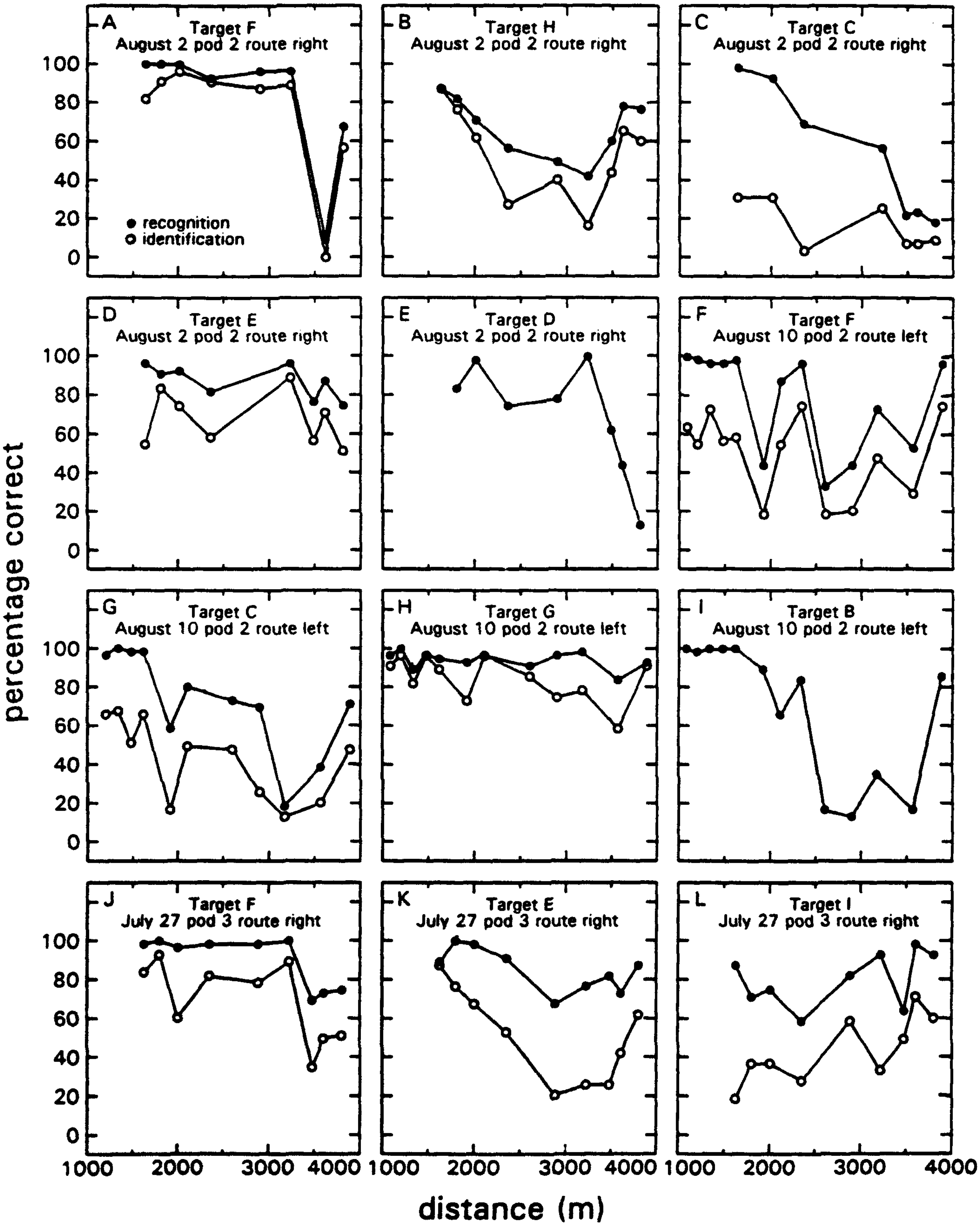
The complete set of observer response data is presented in 126 plots (Figs. A1a-o to A8a-o). All the results are presented as the percentage of correct identification and recognition responses *vs*. target range. The standard deviation of the data points is typically in the order of 10-20%.

Figs. A1-A4 (63 plots) represent the data obtained for the FORCED condition. In this condition (see Chapter 5, section 4), the observers were forced to name the target, even if they were not sure which vehicle was presented. Figs. A5-A8 show the data for the UNFORCED condition. These are the scores for free (unforced) identification or recognition reports, which correspond better to the target acquisition task in a practical military field situation.

Figs. A1-A3 and A5-A7 show the data for the 'position', or 'pop-up', presentation order (indicated as "POS", see section 3.2.4), whereas Figs. A4 and A8 show the data for the "sequential" presentation order (indicated as RUN).
The data were obtained in two series of experimental sessions. In Experiment 1 (Figs. A1, A4, A5 and A8), acquisition performance was determined for 15 daytime runs (POD 2 and 3) of stationary targets (Scenario 1). Both types of presentation order (POS and RUN) were applied. In the figures, the mean scores for 11 observers are presented. In Experiment 2 (Figs. A2, A3, A6 and A7), data was collected for 33 runs of both stationary (Scenario 1) and moving (Scenario 2) targets, on all POD's. Only position presentation (POS) was used. Mean scores are presented for four observers. A complete overview of the structure of the data is presented in Bijl & Valeton (1992c).

FORCED

presentation: pos experiment: 1 scenario:1



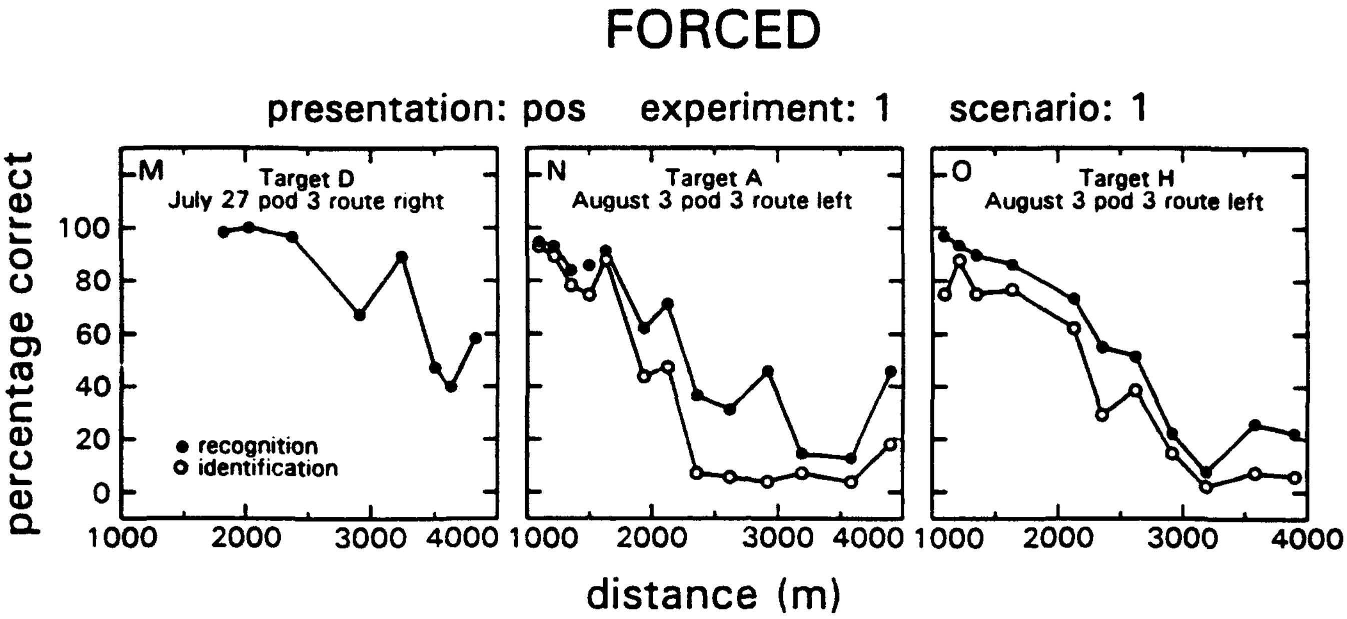
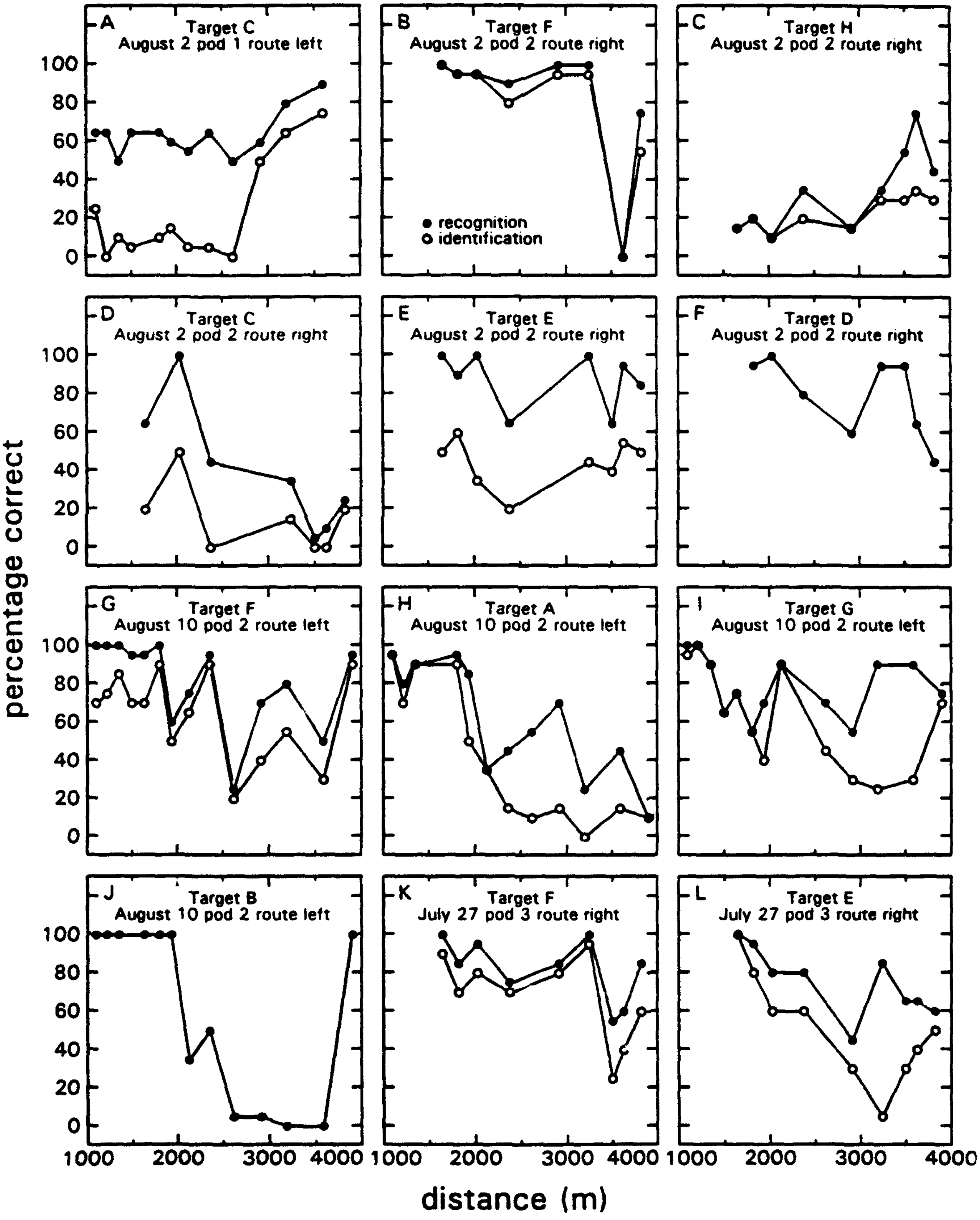


Fig. A1 a-o Observer recognition and identification performance for the condition FORCED-POS-Scenario 1-Experiment 1.

FORCED

presentation: pos experiment: 2 scenario: 1



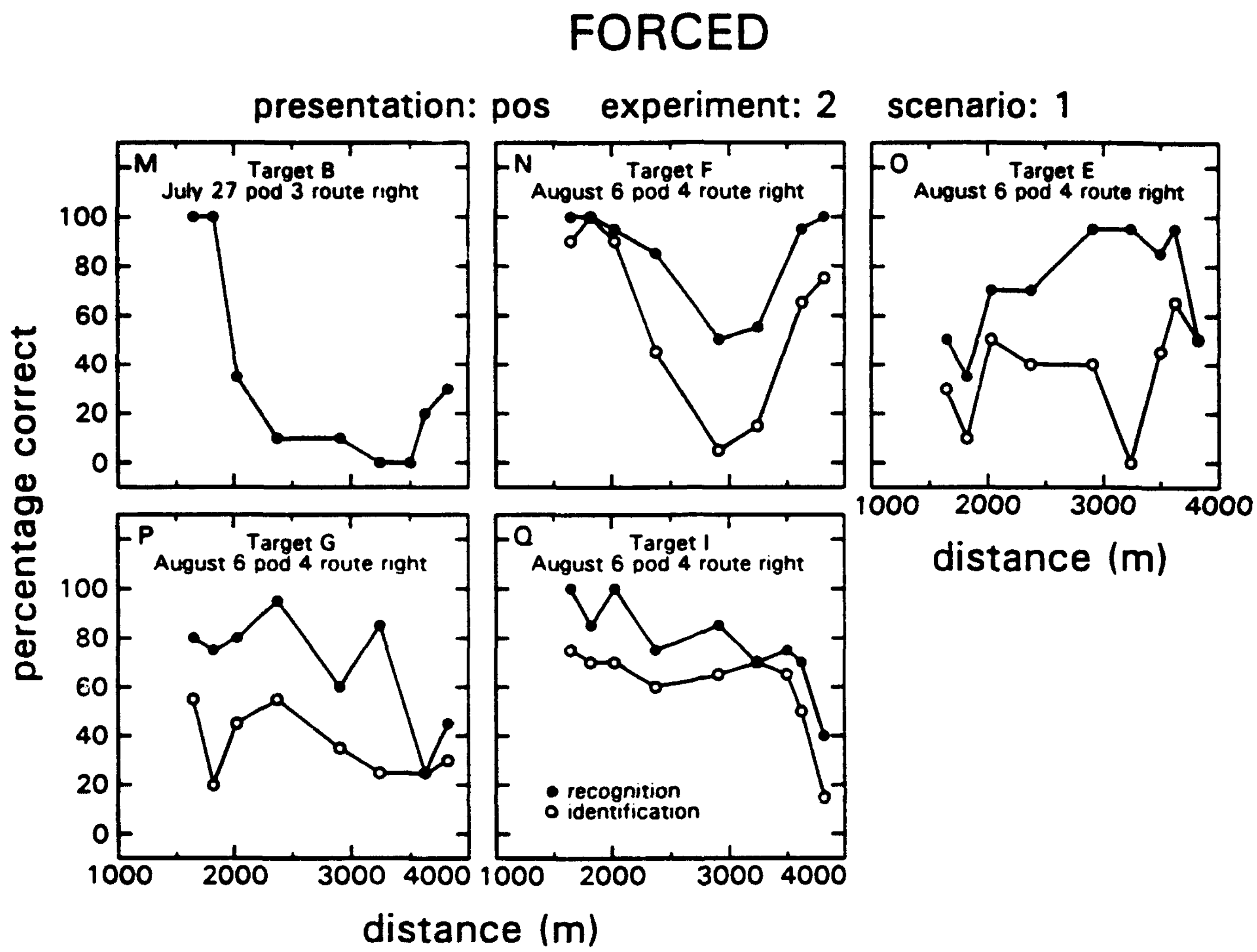
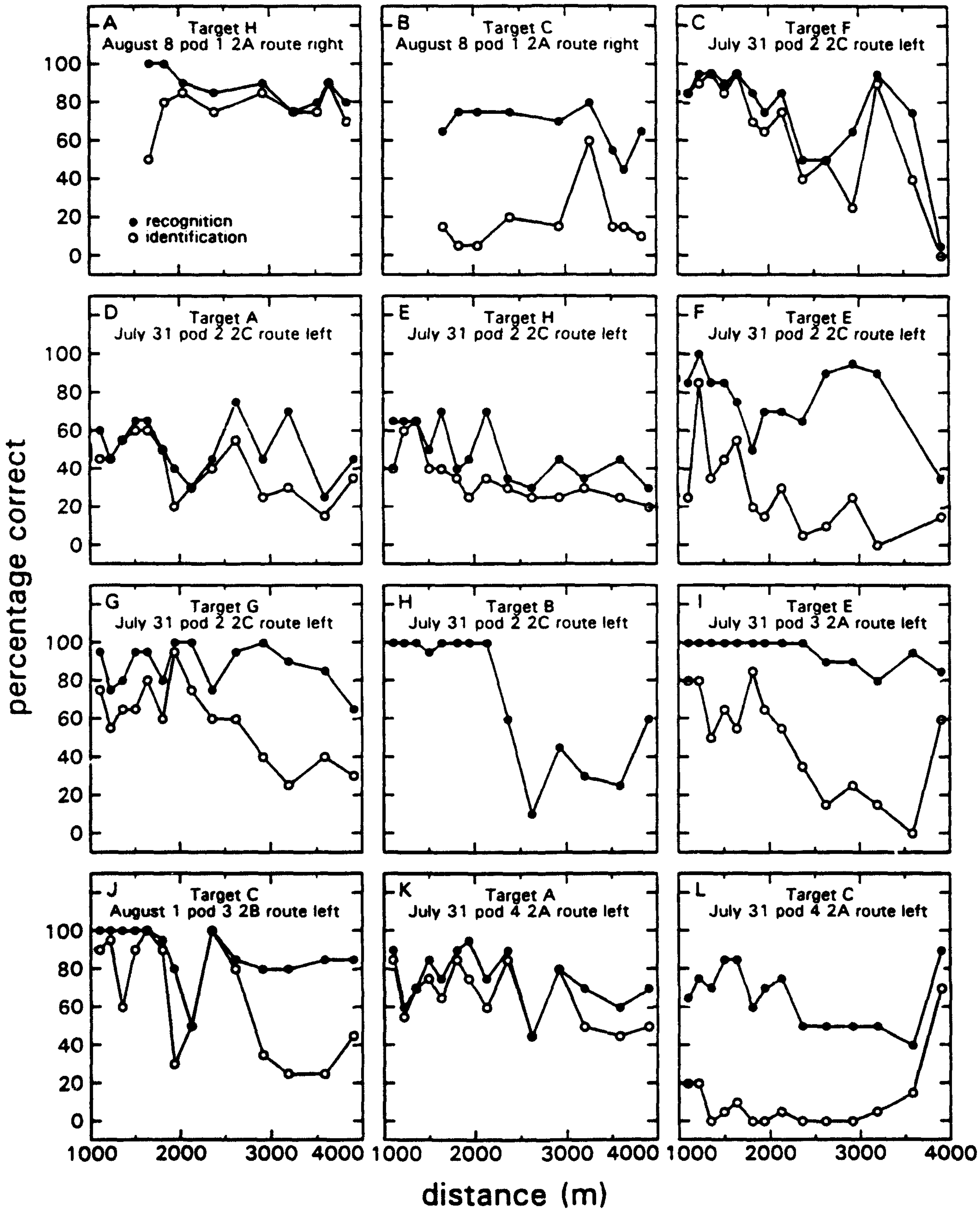


Fig. A2 a-q Observer recognition and identification performance for the condition FORCED-POS-Scenario 1-Experiment 2.

FORCED

presentation: pos experiment: 2 scenario: 2



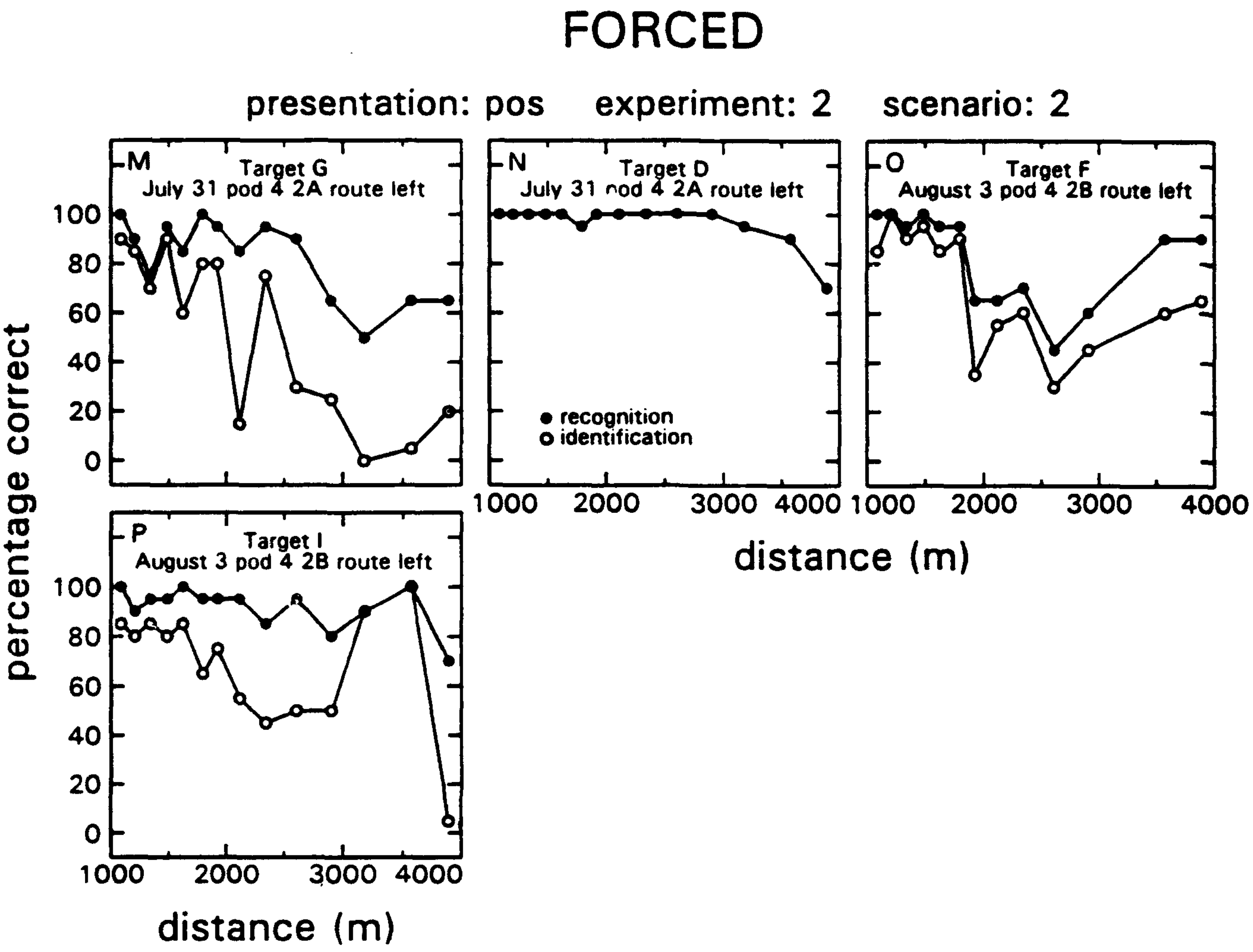
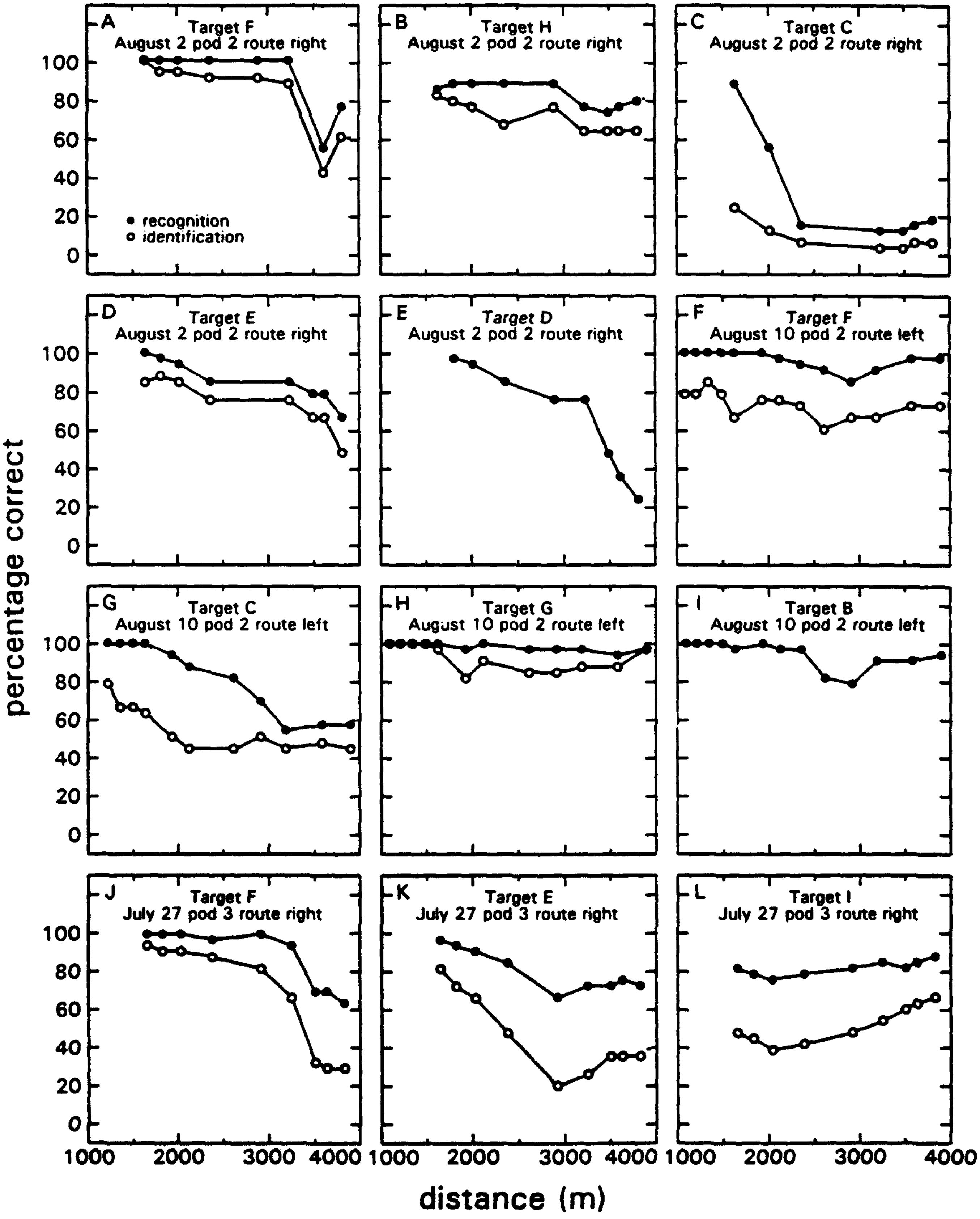


Fig. A3 a-p Observer recognition and identification performance for the condition FORCED-POS-Scenario 2-Experiment 2.

FORCED

presentation: run experiment: 1 scenario: 1



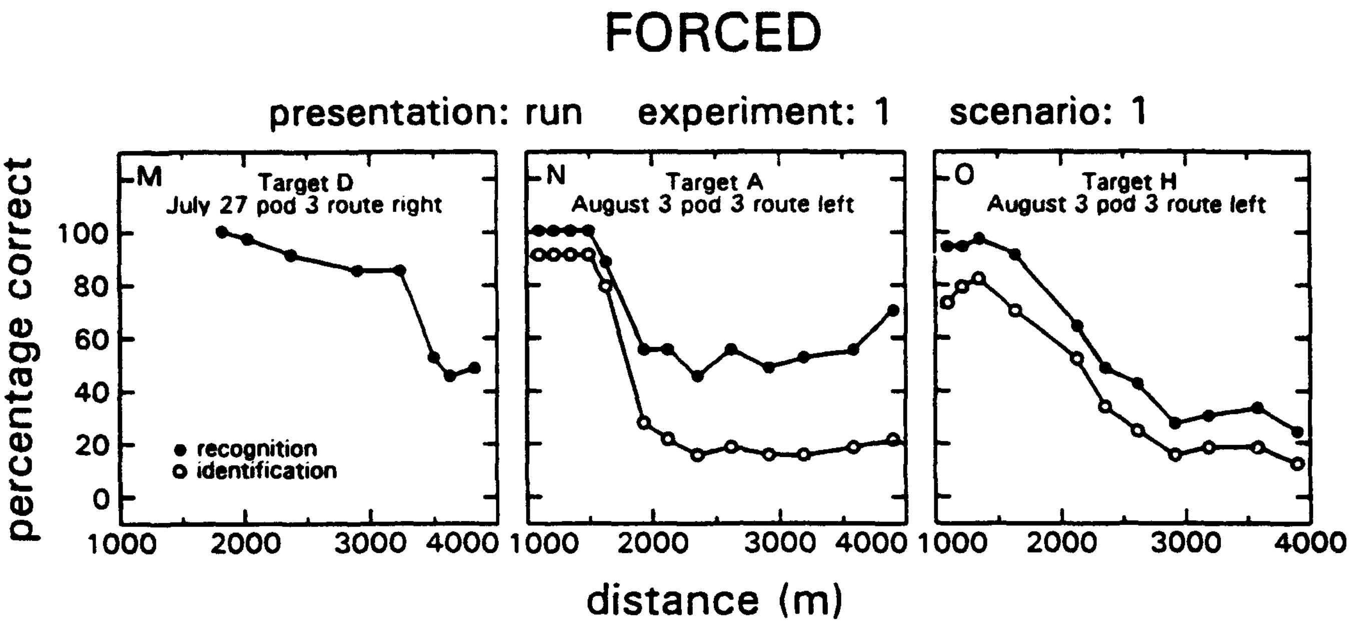
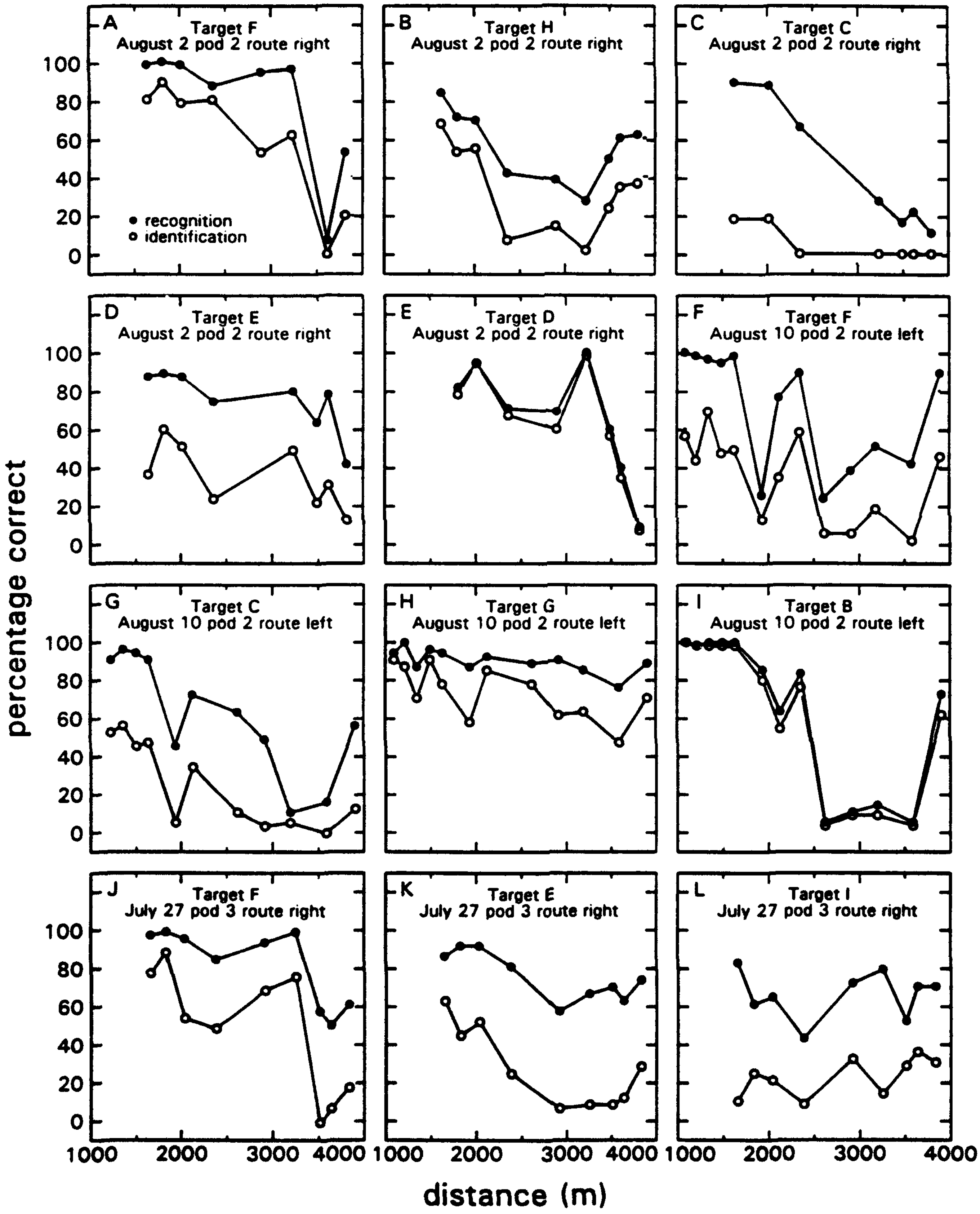


Fig. A4 a-o Observer recognition and identification performance for the condition FORCED-RUN-Scenario 1-Experiment 1.

UNFORCED

presentation: pos experiment: 1 scenario:1



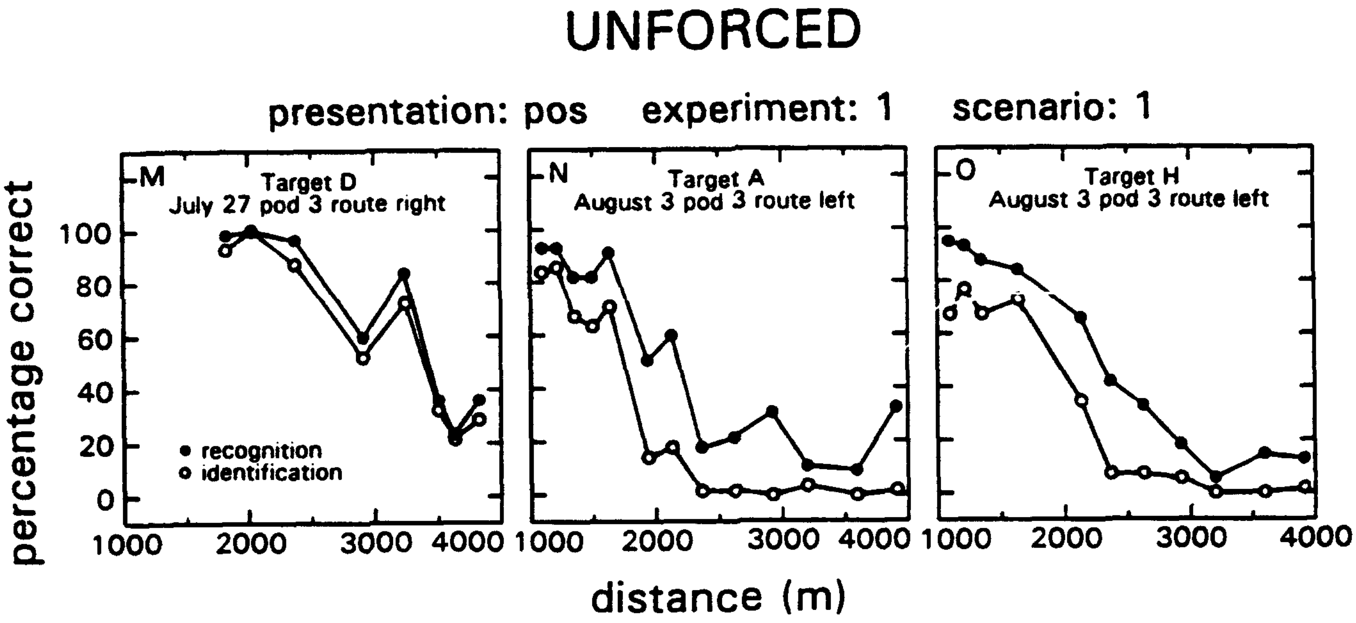
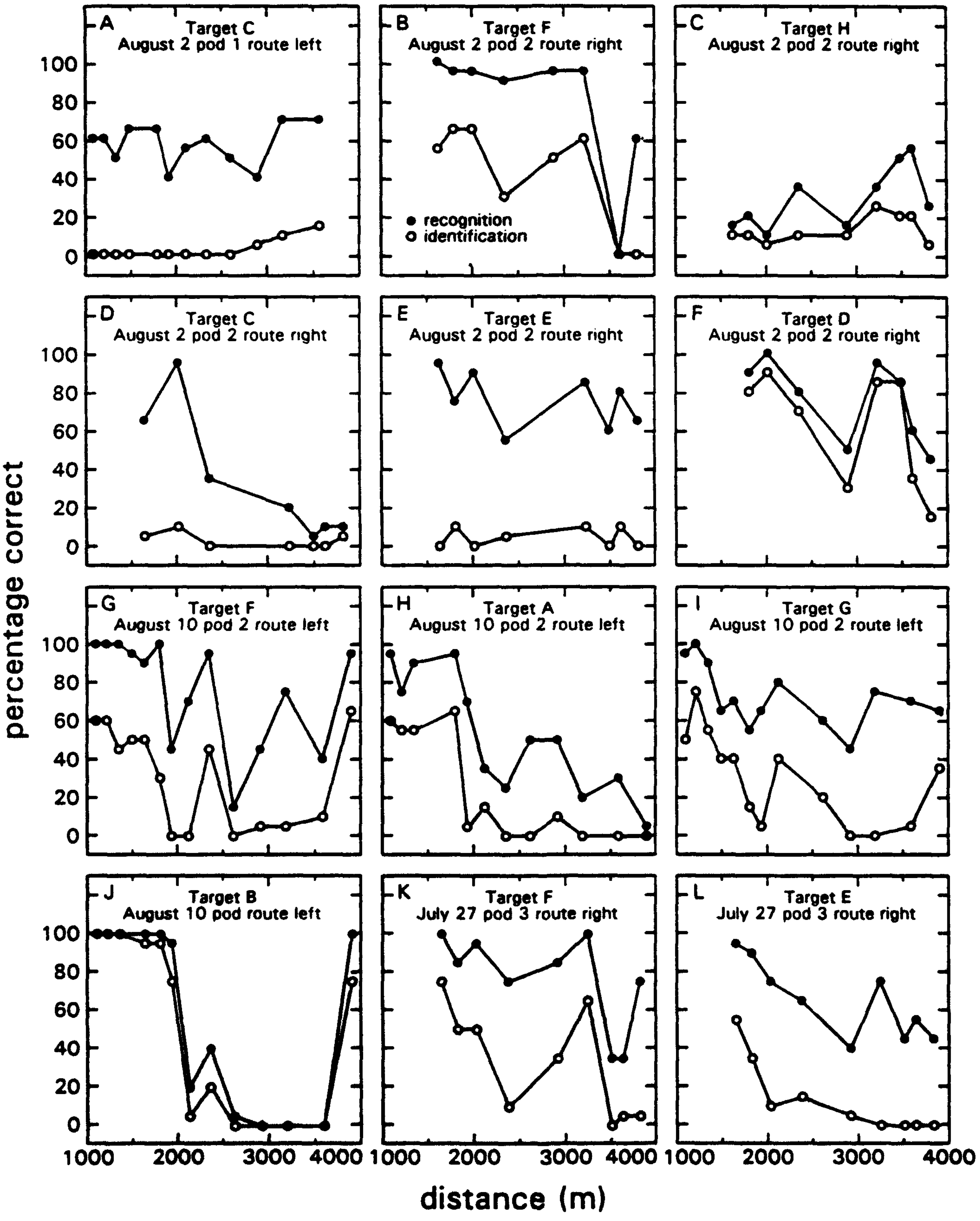


Fig. A5 a-o Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 1-Experiment 1.

UNFORCED

presentation: pos experiment: 2 scenario: 1



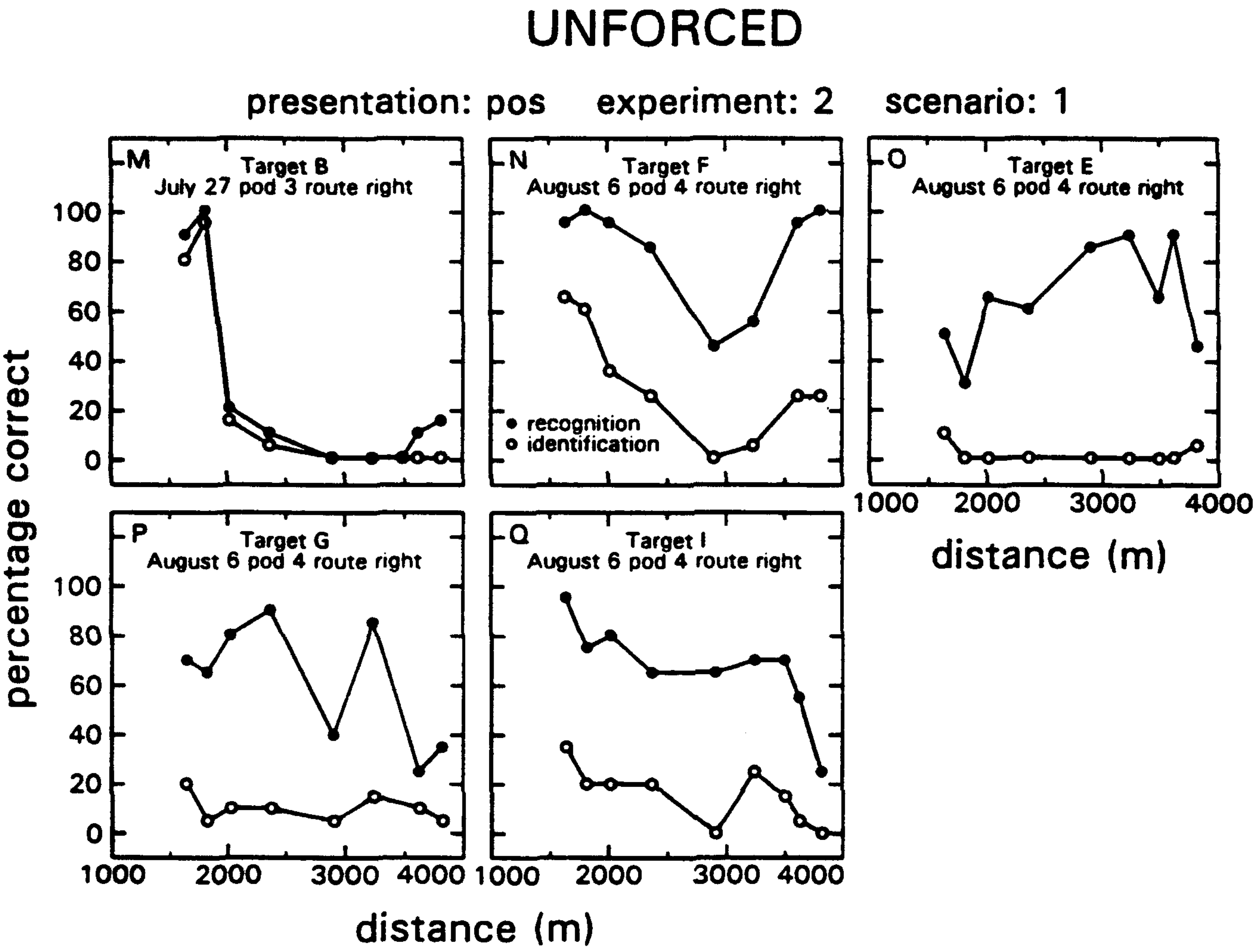
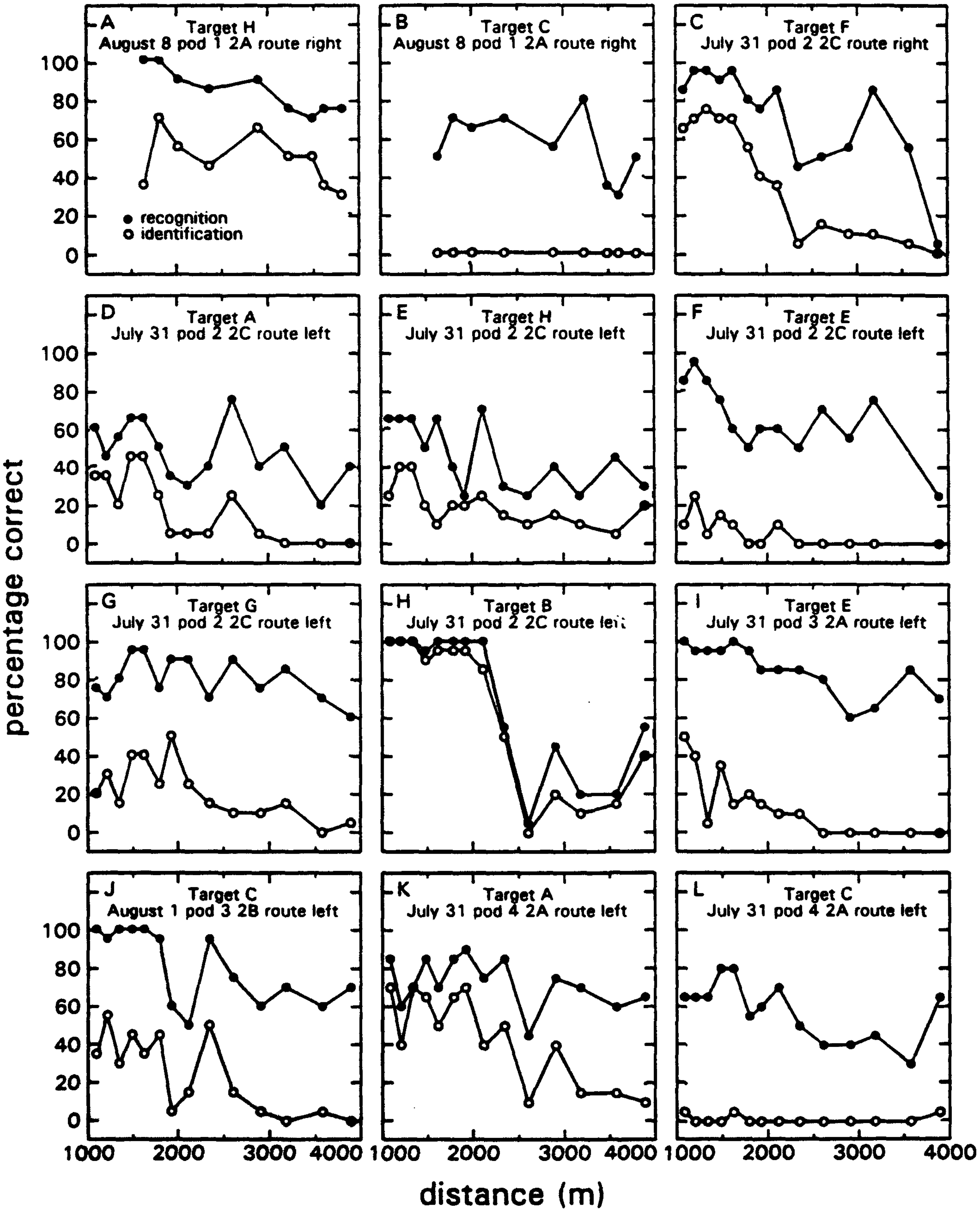


Fig. A6 a-q Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 1-Experiment 2.

UNFORCED

presentation: pos experiment: 2 scenario: 2



UNFORCED

presentation: pos experiment: 2 scenario: 2

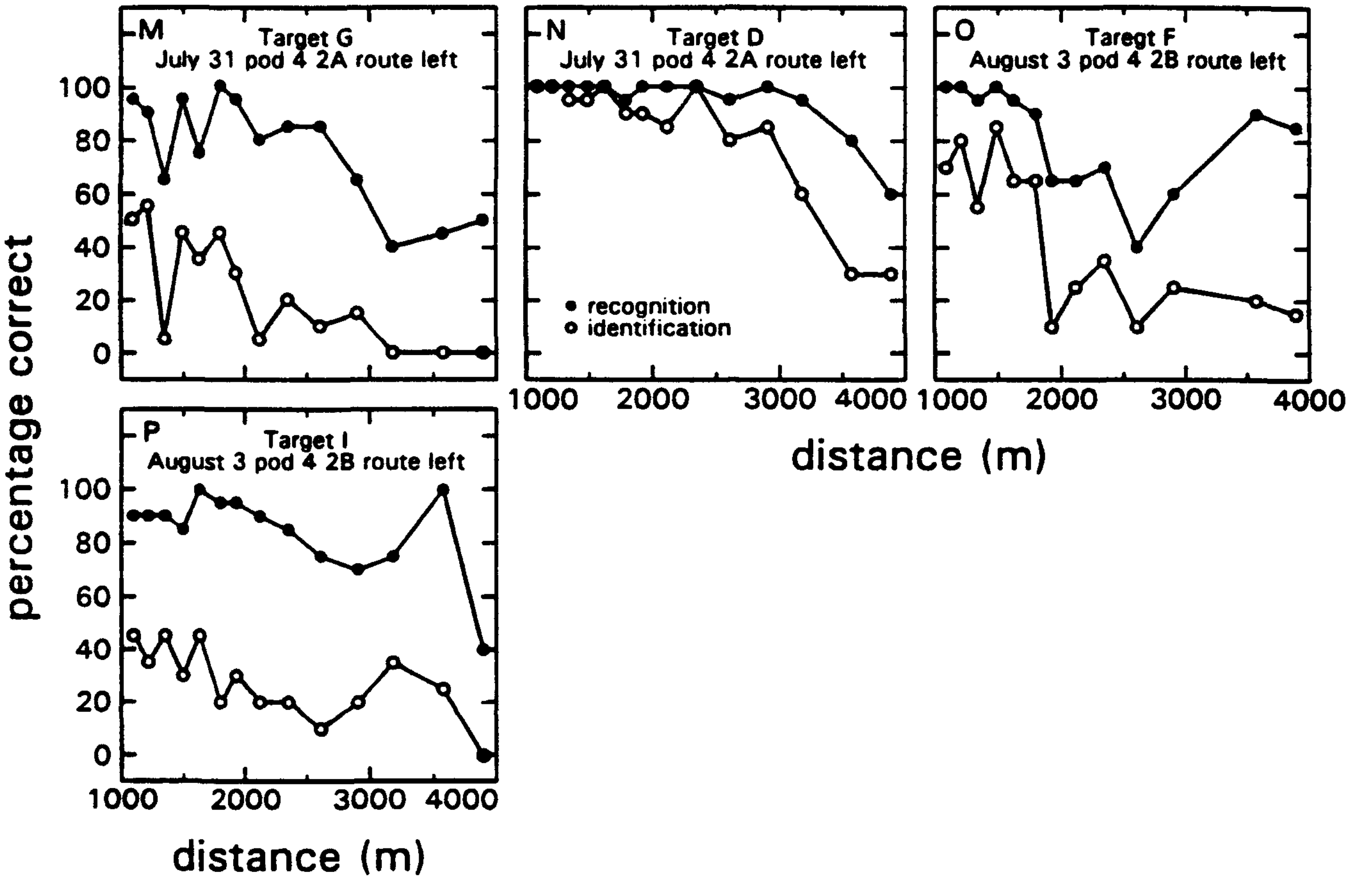
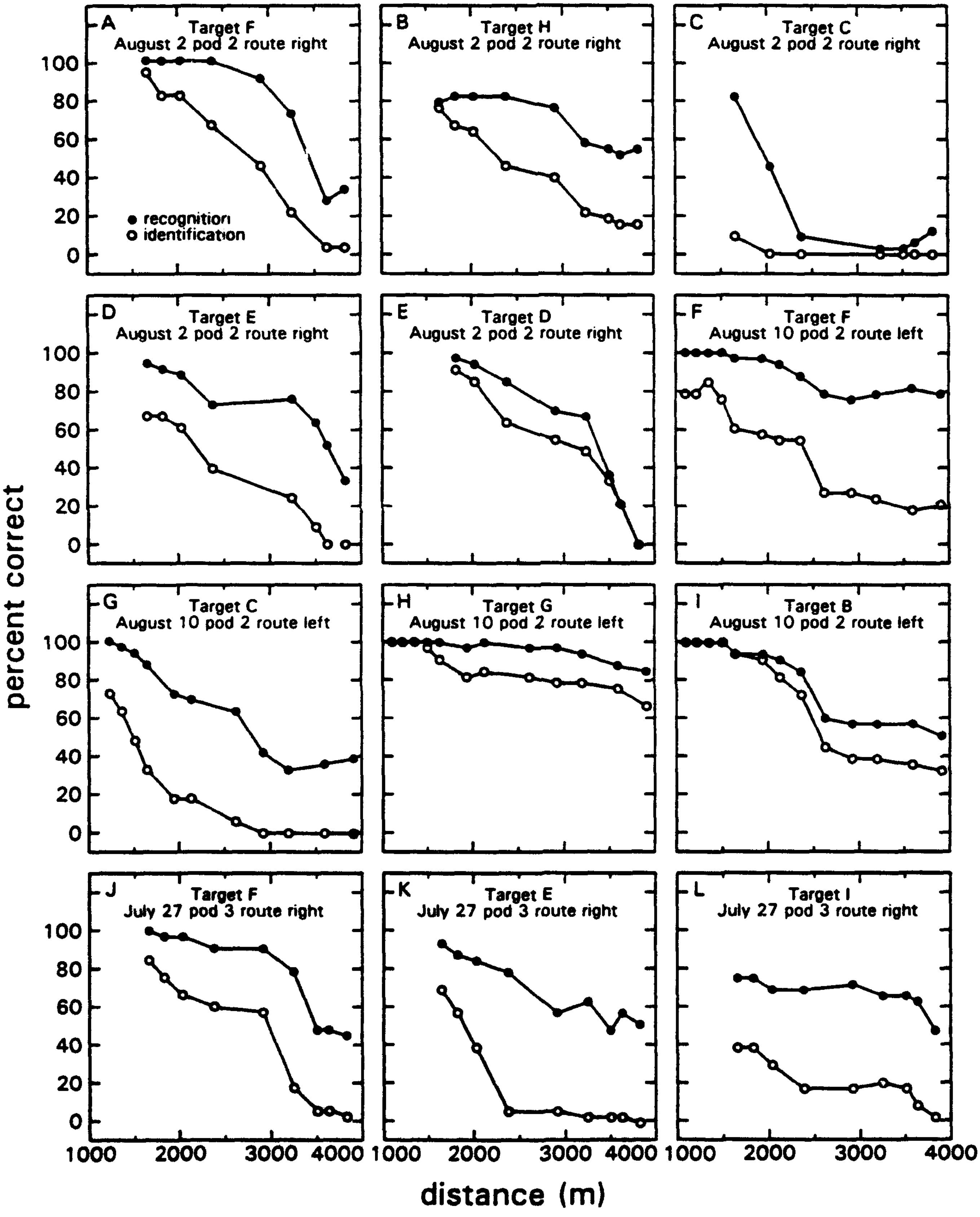


Fig. A7 a-p Observer recognition and identification performance for the condition UNFORCED-POS-Scenario 2-Experiment 2.

UNFORCED

presentation: run experiment: 1 scenario: 1



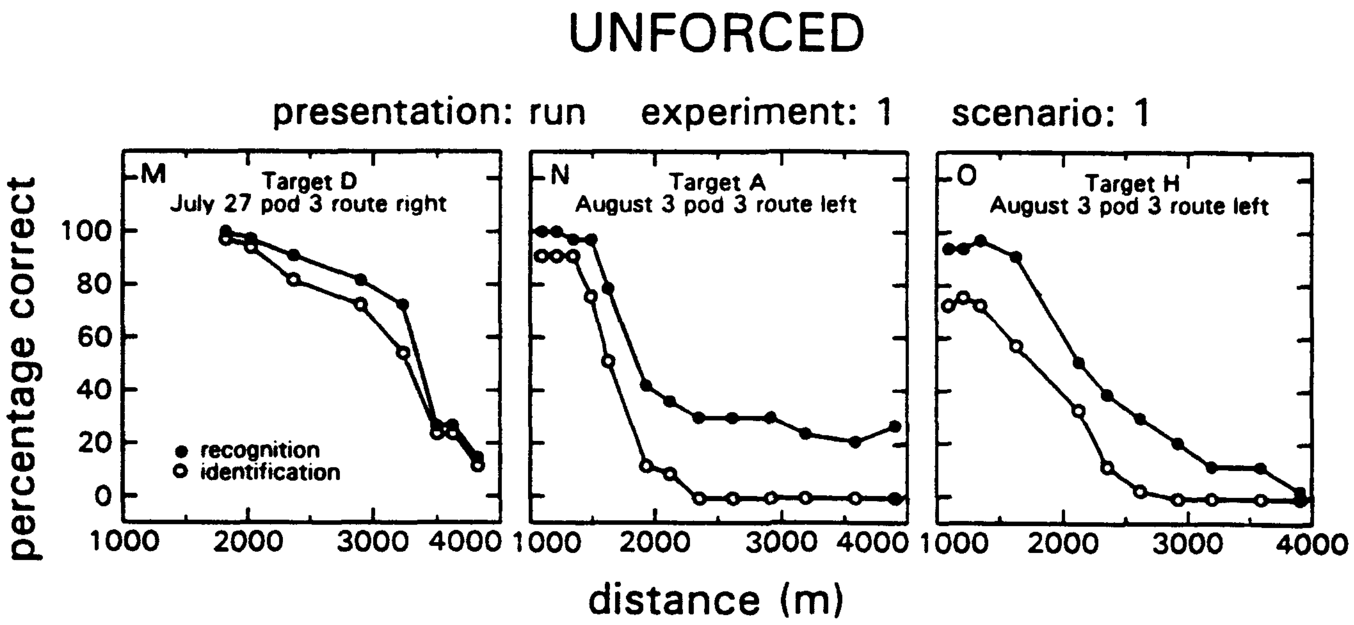
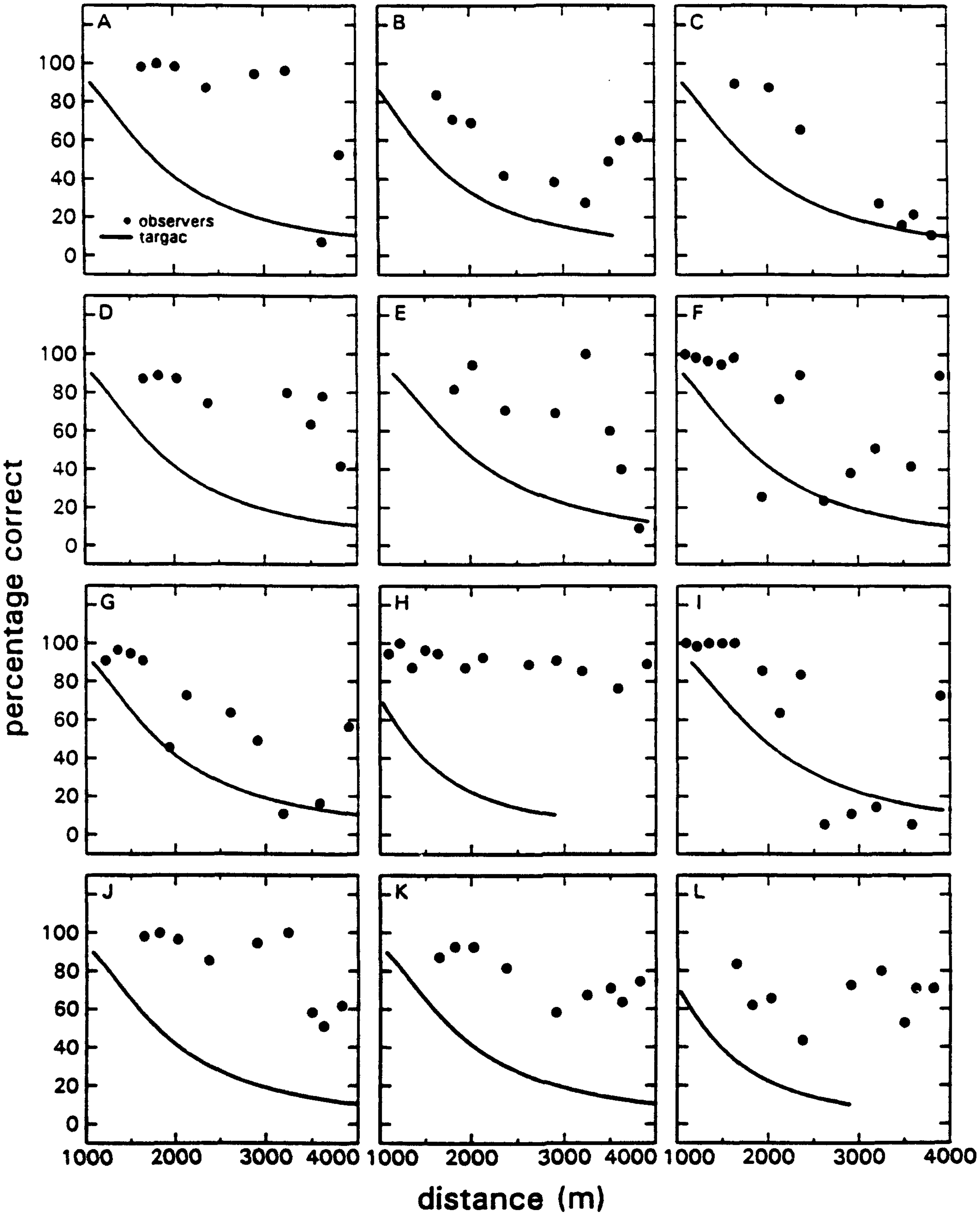


Fig. A8 a-o Observer recognition and identification performance for the condition UNFORCED-RUN-Scenario 1-Experiment 1.

APPENDIX 2: Observer recognition scores and TARGAC predictions for the BEST TWO runs.

The observer recognition scores are taken from Figs. A5-A8 in Appendix 1. These are the scores for free (unforced) recognition reports, which correspond to the target acquisition task in a practical military field situation.

Figs. A9-A10 show the data for the 'position' or 'pop-up' presentation order. Fig. A12 for the 'sequential presentation order.


A
B
C
D
E
F
G
H
I
J
K
L
● observers
targac
100
80
60
40
20
0
1000
2000
3000
4000
percentage correct
distance (m)

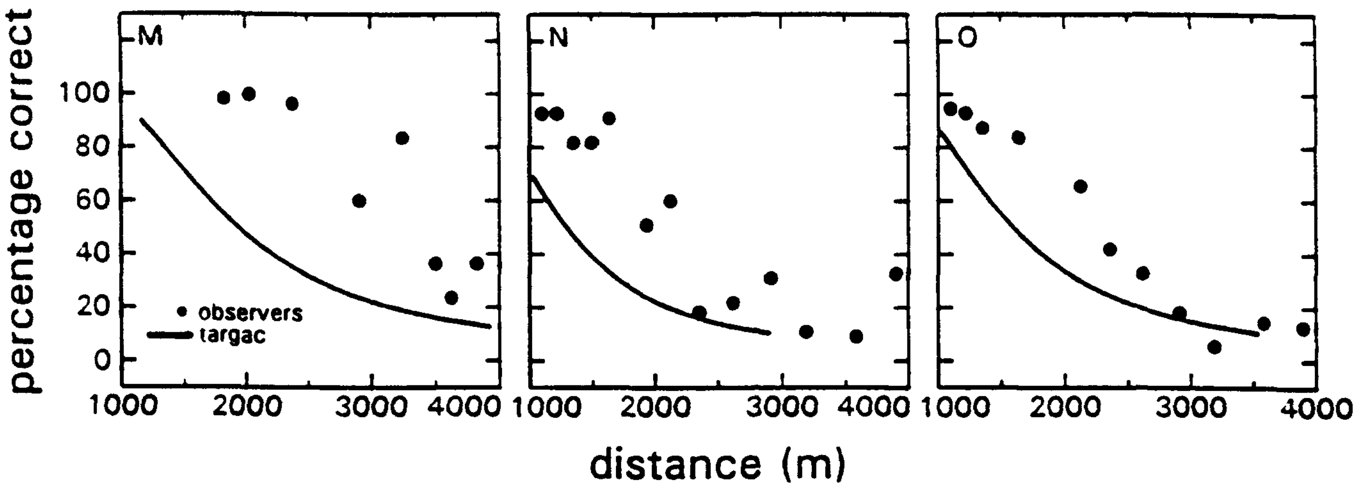
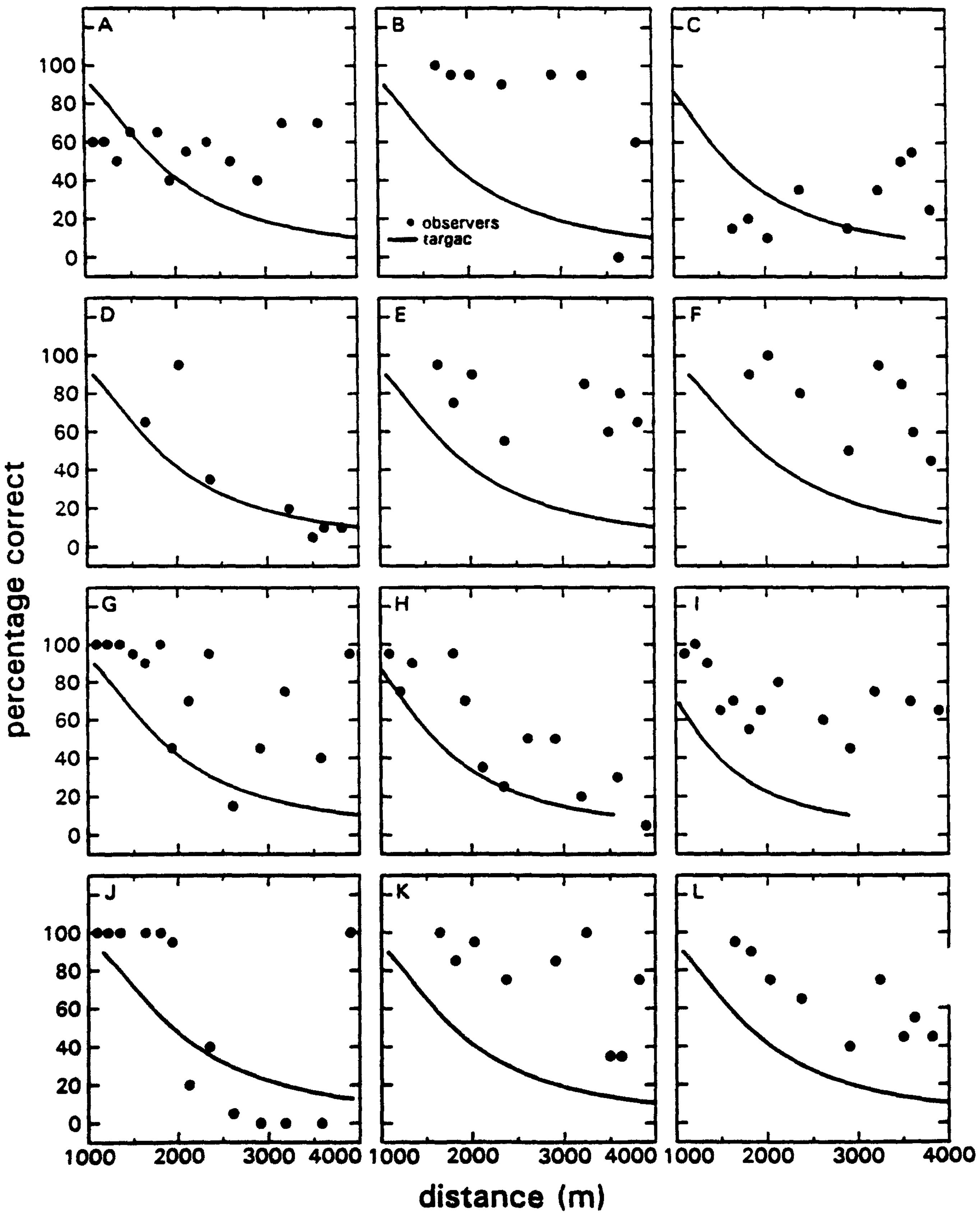


Fig. A9 a-o Observer recognition scores for 15 runs of Experiment 1 (see Chapter 6, section 3.1) for the "pop-up" presentation order, together with TARGAC predictions.


A
B
C
D
E
F
G
H
I
J
K
L
100
80
60
40
20
0
● observers
targac
percentage correct
1000
2000
3000
4000
distance (m)

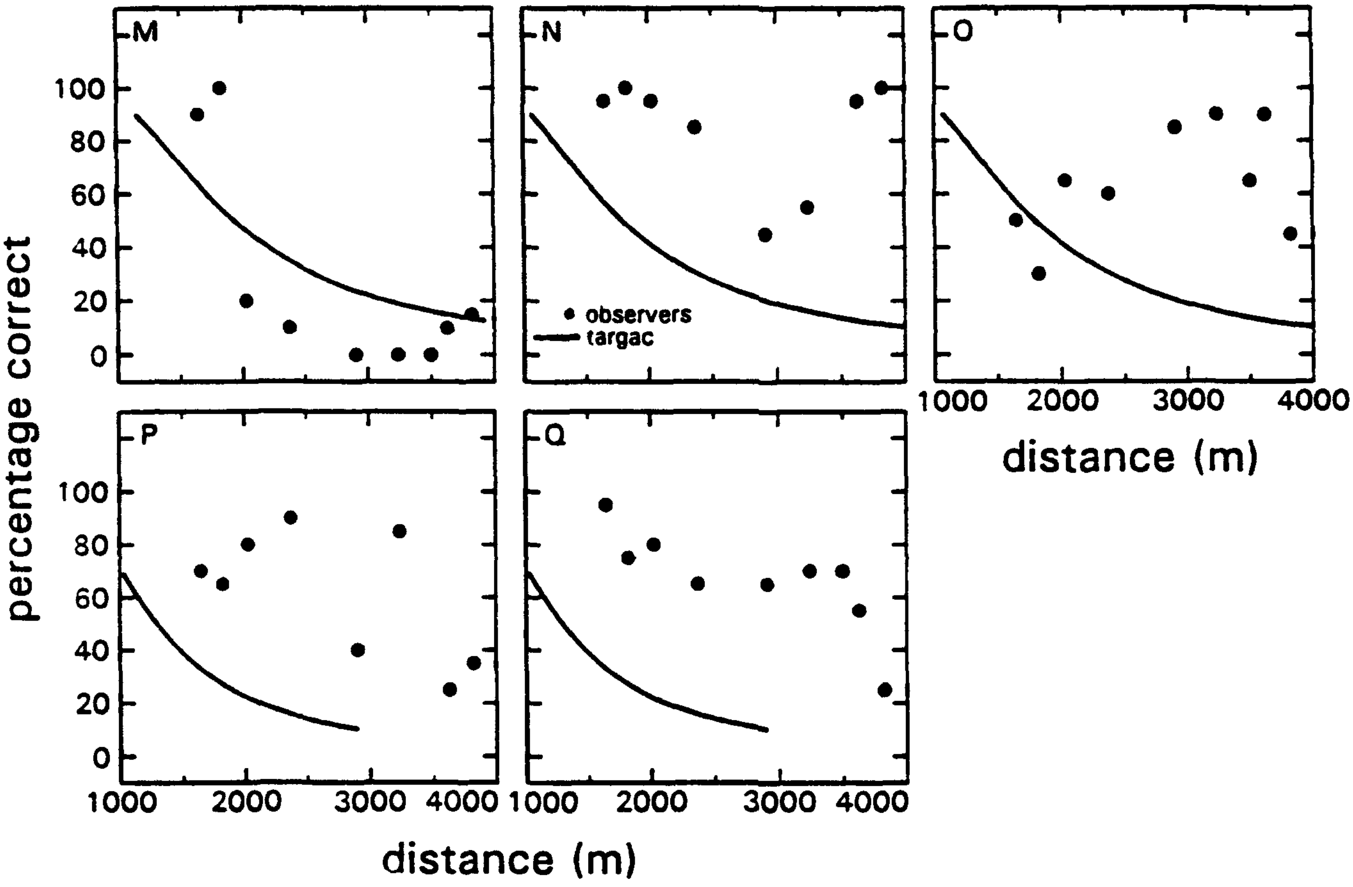
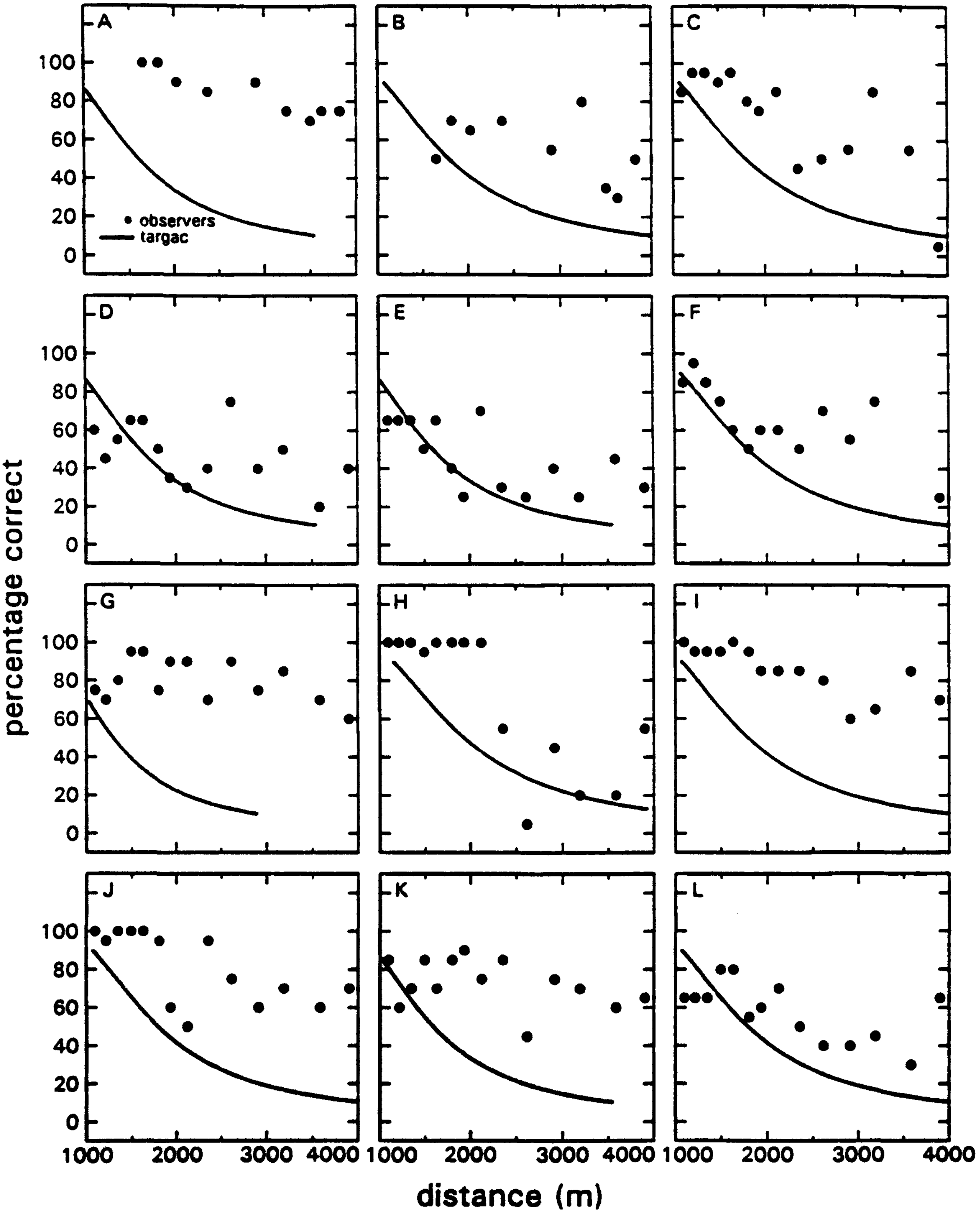


Fig. A10 a-q Observer recognition scores for 17 runs of Experiment 2 (see Chapter 6, section 3.1) for stationary targets, together with TARGAC predictions.


A
B
C
D
E
F
G
H
I
J
K
L
• observers
targac
percentage correct
0
20
40
60
80
100
distance (m)
1000
2000
3000
4000

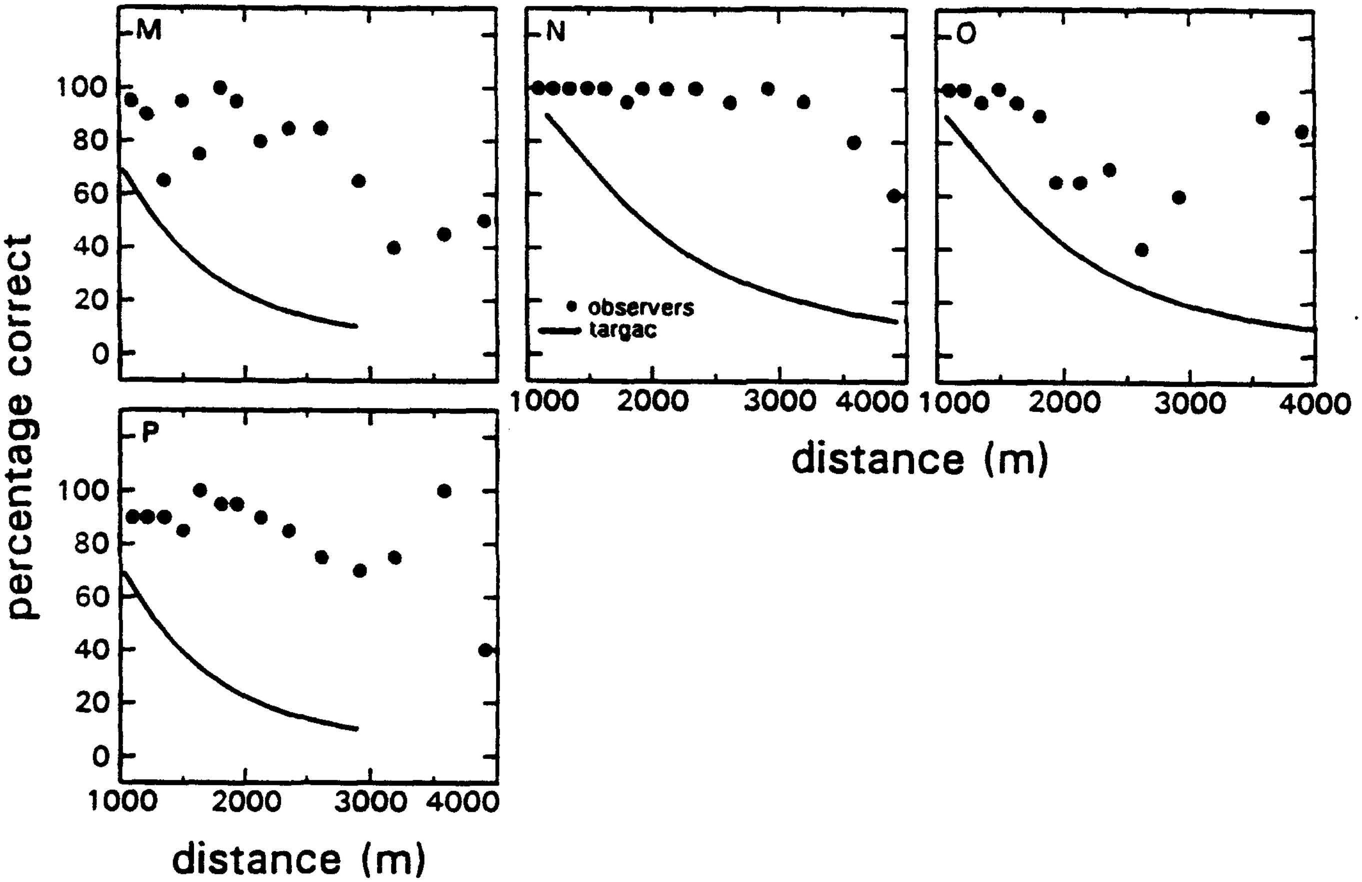
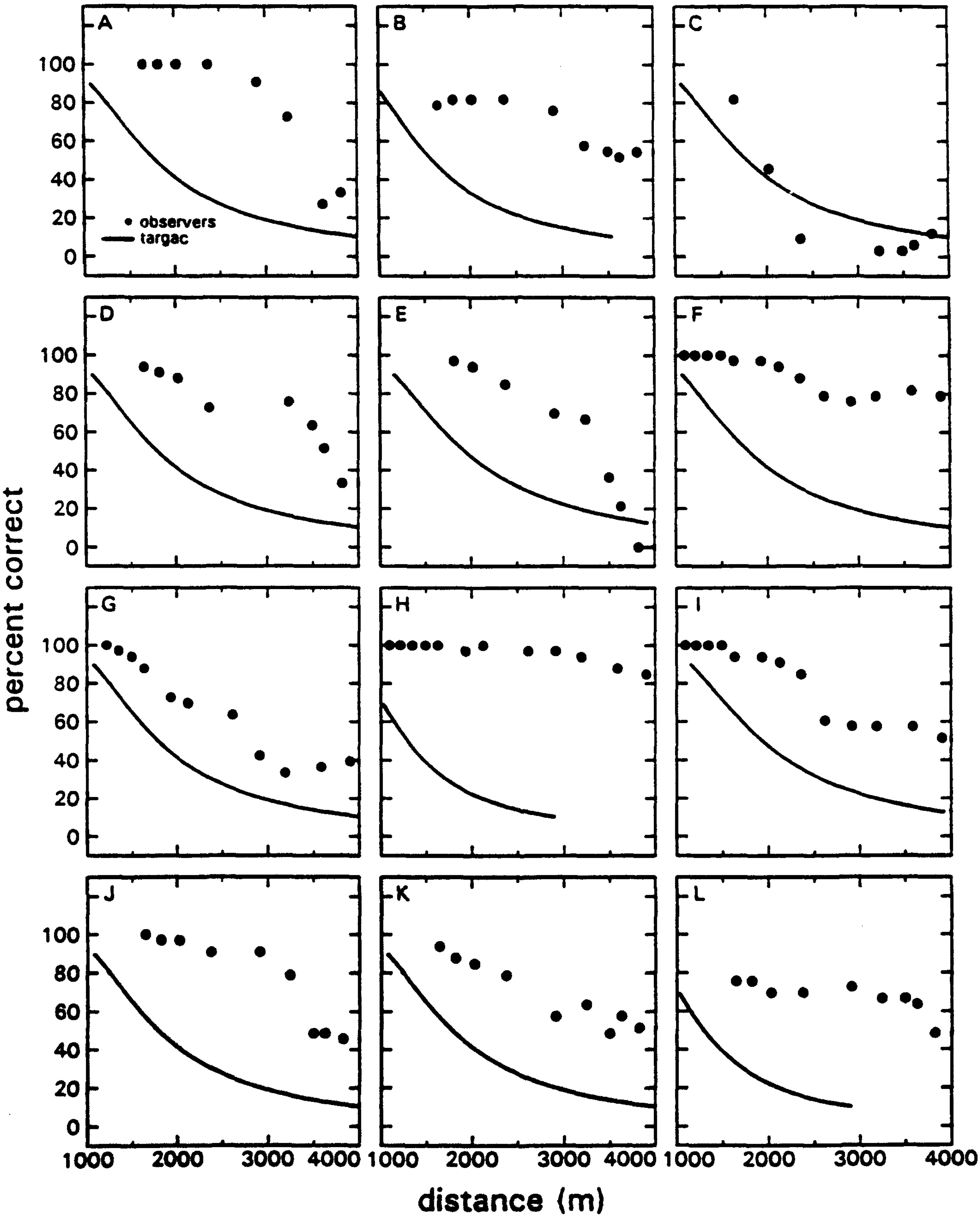


Fig. A11 a-p Observer recognition scores for 16 runs of Experiment 2 (see Chapter 6, section 3.1) for moving targets, together with TARGAC predictions.


A
B
C
D
E
F
G
H
I
J
K
L
• observers
targac
percent correct
0
20
40
60
80
100
1000
2000
3000
4000
distance (m)

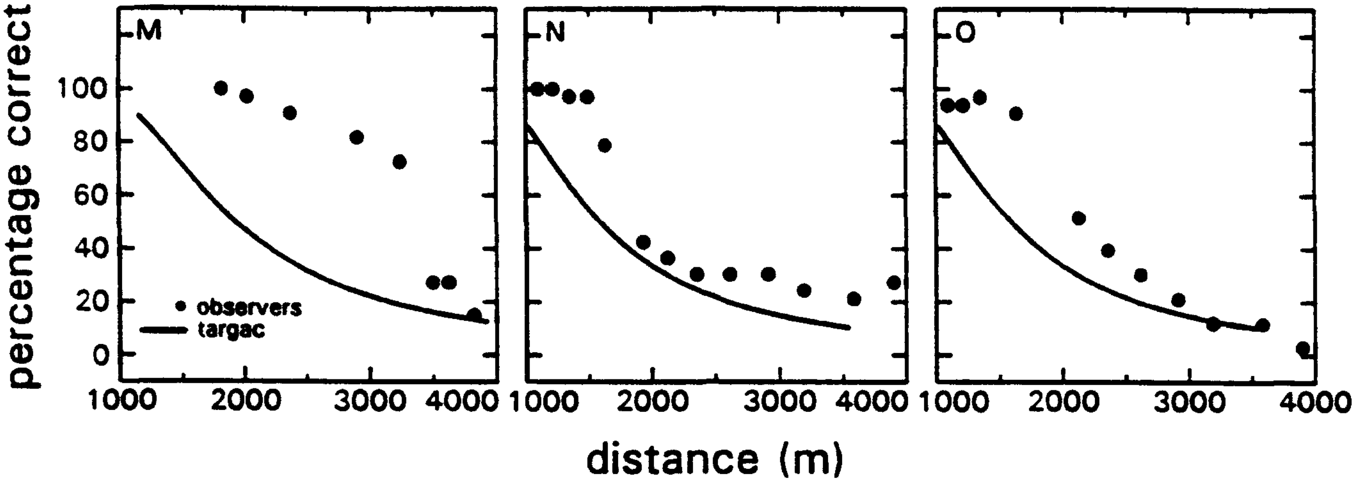


Fig. A12 a-o Observer recognition scores for 15 runs of Experiment 1 (see Chapter 6, section 3.1) for the "approaching" presentation order, together with TARGAC predictions.

APPENDIX 3: Statistical errors in the observer scores, and their effect on *(r/r')*.

A: Estimate of the error in the observer scores

The observer performance data were gathered in two experiments (see 3.1). In Experiment 1, the number of observers was 11; Four observers participated in Experiment 2. The images of ten target runs were used in both experiments, thus being presented to 15 observers. Data set A was composed of observer performance data from both Experiment 1 and 2 (which means that the number of observers is 4 - 15). Data set B and C were composed of data from Experiment 1 only (11 observers). Each target image was presented 5 times. In Chapter 3 it is shown that, if the scores are distributed binomially, the maximum standard deviation s_{max} (being the standard deviation at the 50% probability level) is 7-15% for 11 observers, and 11-25% for 4 observers. The lower estimates of s_{max} are based on the assumptions that all observations are independent; for the higher values a worst-case assumption was made that all the 5 observations of the same image by the same observer are completely dependent.

The standard deviation at probability levels above or below 50% is given by:

(A1) $$\sigma_p = \sigma_{max} \frac{\sqrt{P\ (100-P)}}{50}$$

where P is the probability level in percent. From equation A1, it can be deduced that the standard deviation is rather constant at levels between 20% and 80% and drops to zero if the probability is near 0 or 100%.

The above-mentioned estimates of σ_{max} are not very accurate. The standard deviation can be estimated in an alternative way by dividing the observers into two groups, and calculating the correlation between the scores of the two groups. If the correlation coefficient r is high, the statistical error in the scores is small. It can be shown that an accurate approximation of the error variance (this is the variance in the scores for each image, due to a limited number of observations), is given by:

(A2) $$\sigma_e^2 \approx \frac{(\sigma_{G1}^2 + \sigma_{G2}^2)\ (1-r)}{4}$$

where σ_e^2 is the error variance in the score for each image, averaged over all observations of all observers of the two groups, and σ^2_{G1} and σ^2_{G2} are the total variance in the scores for the images, when averaged over the observations of the observers within each group.

The correlation was calculated for the images of the ten runs that were presented to 15 observers. Observers from the two experiments were divided equally over the two groups. The correlation is very high: if all data (96 data points) are taken into account, r = 0.92, yielding a standard deviation σ_e = 4.1%. However, this is an average over all probability levels between 0 and 100%. The standard deviation for low and high scores is much lower than for intermediate probability levels. As indicated above, for probability levels between 20% and 80%, the standard deviation may be regarded as rather constant. If we take only these data (38 data points) into account, we have an approximation for the maximum standard deviation σ_{max} (at the 50% level). It turns out that for these data the correlation is still very high: r = 0.86, and σ_e = 5.0%. This is the estimate of the maximum standard deviation for 15 observers. As the standard deviation is inversely proportional to the square-root of the number of observers, the maximum standard deviation in the data in set A (4 - 15 observers) is σ_{max} = 5 - 10%, and in set B and C (11 observers) is σ_{max} = 6%. These values are comparable to the lower estimates based on a binomial distribution.

B: Contribution to the error in the (r/r')-values

For each trial, actual target range *r* is known with high accuracy. The predicted range, *r'*, is calculated from the value of the recognition probability *P* by using TARGAC, or equation 4 in section 5.1. An error σ_P in probability will result in an error $\sigma_{r',obs}$ in range or $\sigma_{(r/r'),obs}$ in the ratio. The ratio $(\sigma_P/\sigma_{r',obs})$ is given by the slope of the probability *vs.* range function (equation 3). For example, at the 50% level the slope of the function (with s = 2.32) is -0.80/r_{50}. Because the function is shallow at very high or low levels, a small error in recognition probability may lead to a large error in range. Between 20% and 80%, the standard deviation is approximately constant, and the probability *vs.* range function is roughly linear. It is therefore safe to consider only data between these probability levels. It can be shown that, for these data, the standard deviation in ratio, due to the error in observer scores, is given by:

$$\sigma_{(\frac{r}{r'}),obs} \approx 1.5\ \sigma_{max} \qquad \text{(A3)}$$

Thus, for data set A, $\sigma_{(r/r'),obs} \approx$ 8 - 15% (0.03 - 0.06 on a log scale), and for sets B and C, $\sigma_{(r/r'),obs} \approx$ 9% (0.04 on a log scale).

APPENDIX 4: List of errors in TARGAC software found during evaluation

During the evaluation of TARGAC, the following software errors were found:

1. *Conversion errors*

For a single variable, several units are used in the program. For example, size or distance are expressed in meters, kilometers or feet. It occurs in a number of subroutines, that *global* variables, the value of which is written in a common block, are converted from one unit to another. The new value of the variable will be used in the next routines (through the common block), although a number of those routines expect the variable in the old unit. Incidentally, the same conversion is carried out twice (this occurs with the conversion of the rain rate in mm/hour to inch/hour). It is recommended that the value or unit of global variables is never converted. If the value of a variable is desired in a different unit, a local variable should be defined.

2. *Sensor altitude*

In TARGAC, both target and sensor height may be varied. A supplement of the User Guide reports, however, that the slant path option (looking down to a target), does not work correctly, and that the sensor altitude is currently hardwired to 0 or 1 m. What actually happens in the subroutine that calculates acquisition ranges (FINDR), is that sensor altitude is temporarily set to 1, and later on to 0. In this routine, however, ranges are defined in km! Thus, the program calculates ranges for a sensor looking down from a height of 1 km, whereas the output file gives an altitude of 0. This error has important consequences for the range predictions: for a target in front view, the minimum or effective dimension is target height. However, looking down from an altitude of 1 km, its effective dimension is target width, which is usually larger. The error was repaired by setting the altitude to 0 in FINDR.

3. *Extrapolation of a high order polynomial curve fit*

When predictions are being made for a user-specified viewing device (see section 2.3), the user has to specify spatial frequency as a function of (luminance or thermal) contrast in the form of the coefficients of a sixth order polynomial fit rather than a point-by-point entry of the MRC/MRTD curve (actually, point-by-point entry is another TARGAC option to specify a viewing device, but it does not work properly). Polynomial curve fits may only be used for interpolation: extrapolation of a high order polynomial function may lead to irrational results. However, thermal contrasts, as calculated by TCM2, are often much higher than the highest contrast in the MRTD curve. As a consequence, unrealistic spatial frequency values (very high values, and even negative values occur!) are calculated, which in return lead to meaningless range predictions. No warning is given to the user. The problem probably also occurs with viewing devices that are in the TARGAC menu (Gillespie, 1993). We recommend the following improvements. A: If the apparent contrast exceeds a given limit (e.g. the highest contrast of the MRC/MRTD curve), a warning should be given (actually, a lower contrast limit already exists in TARGAC). B: If the limit is exceeded, calculations may be carried out using the spatial frequency that corresponds to the contrast limit. C: A lower order polynomial fit should be used: a second or third order fit should be sufficient. In the present evaluation, the problem was circumvented by adding extra MRTD points for thermal contrasts up to 20 K before making a polynomial fit.

4. *History of meteorological data*
TCM2 calculates target and background temperature at a given moment on the basis of meteorological data for a number of earlier moments in time, e.g. 0, 3 and 6 hours earlier. These data are treated differently in interactive and batch mode. The user guide is correct only for the interactive mode. In batch mode, history is reversed if the input file is constructed according to the user guide: a correct calculation is made if the preceding times are given as negative numbers. Accordingly, files saved in the interactive mode, and input files for the batch mode, are incompatible with respect to this point.

5. *Target heading*
In the input, one is free to choose the heading of the target. In the program, however, target heading is always set at 90 degrees.

6. *Wrong target files*
For several targets in the menu (targets 19 through 22), TCM2 rendered a temperature of 0 K. Range calculations were made using this target temperature, and no warning was given to the user. Later, new target files were provided by Dr. Gillespie that gave reasonable temperature values.

7. *Version number and release date*
Regularly, new versions of the program have been released. However, the version number and release date are not updated.

8. *Undeclared variables*
A number of variables that are used in the program, are not declared.